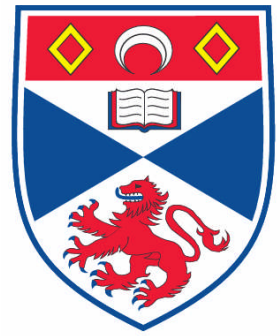
USING GENERALIZED ESTIMATING EQUATIONS WITH REGRESSION SPLINES TO IMPROVE ANALYSIS OF BUTTERFLY TRANSECT DATA

Ciara Brewer

A Thesis Submitted for the Degree of MPhil
at the
University of St. Andrews



2008

Using Generalized Estimating Equations with Regression Splines to improve Analysis of Butterfly Transect Data

Ciara Brewer

A Thesis submitted for the degree of M. Phil.

August 2007

Abstract

Surveying animal populations is an important aspect of wildlife management. Distinguishing trend from random fluctuations and quantifying trend are key goals in any analysis.

The aim of this thesis is to review analyses of Butterfly Monitoring Survey (BMS) data and to develop new methods which address some flaws in previous studies. The BMS was established in 1976 at Monks Wood, Cambridgeshire and sites were added over time throughout Britain in order to monitor butterfly population trends. Weekly counts are made over the monitoring season and the main aims are to produce annual indices and compare these indices over time for any particular species.

Originally, weekly counts were summed to produce relative indices and missing counts were estimated using linear interpolation. This thesis discusses the weaknesses of this basic method and suggests possible improvements.

In recent years, with advancements in statistical methods and increased computer power, new methods can be applied to accommodate the longitudinal and flexible nature of ecological data.

Mixed Models, Generalized Estimating Equations and Generalized Additive Models are used and the relative merits of each modelling approach discussed. These methods allow for correlation and non-linearity in data.

Model selection is an important consideration when modelling and different tests are introduced and compared.

Once a model is selected, site-level indices are estimated, which can be col-

lated to produce regional and national indices. Different methods of estimating precision around indices are also contrasted. Bootstrapping is found to be a convenient and dependable approach.

Abundance is difficult to disentangle from detectability when only counts of species are carried out. Methods for dealing with this problem are suggested.

Once reliable annual abundance estimates are found, they can be compared over time using a variety of statistical techniques. The chain-ratio method is applied to a subset of real data.

Declarations

I, Ciara Brewer, hereby certify that this thesis, which is approximately 20,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date..... signature of candidate.....

I was admitted as a research student in [July, 2003] and as a candidate for the degree of M. Phil. in [July, 2003]; the higher study for which this is a record was carried out in the University of St Andrews between [2003] and [2007].

date..... signature of candidate.....

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of M. Phil. in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date..... signature of supervisor.....

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker.

date..... signature of candidate.....

Acknowledgements

Thank you to my supervisor, Steve Buckland, without whose help I would never have completed this thesis.

Thanks to my parents for supporting me throughout my time at St Andrews.

Thanks to everybody in the Statistics department, especially Peter Jupp for his encouragement and Charles Paxton for sharing his computers.

Thanks to Rhona and Phil for keeping things running smoothly.

Thanks to Jon and Tiago for answering my many stupid questions.

Thanks to Tom Brereton for use of data, and to all the volunteers who collected data over time.

Thanks to Malcolm and Barbara for helping me to keep sane.

Thanks to all my friends who have supported me:

In Ireland: Katie, Clár, Eimear, Danny, Claire, Róisín, Sinéad, Mary, Jean, Lloyd, Dervla

In Scotland: Blerina, Graeme, Tad, Rachel, Julia, Elizabeth, Kristin, Kristin, Helen, Sarah, Joseph, Laura, Gil, Jenni, Kieran, Eleanor, Stewart, Marco, Kenny, Lucia, Paul, Glee

From 46 RR: Rachel, Claire, Paul, Ben, Dan

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis is to develop and improve methodologies for analysing butterfly monitoring scheme (BMS) survey data collected in the United Kingdom since 1977 and to review different methods of analysis.

During the 1960's, concern was growing over the use of organic chemicals in farming and in the wider countryside, and the effect of these chemicals on the environment. Many surveys worldwide of many different taxa were established at this time in order to investigate abundances and trends (both temporal and spatial) in different species. Amongst these surveys are the UK Common Bird Census (UKCBC) and the North American Breeding Bird Survey (NABBS). The Butterfly Monitoring Scheme (BMS) officially began in 1976, after preliminary counts for 3 years on sites at Monks Wood, Cambridgeshire. The Centre for Ecology and Hydrology (CEH) established the BMS in order to monitor the number of butterflies occupying selected sites throughout Britain and to use these abundances in order to assess trends in butterfly populations over time. Sites used were mainly national parks and nature reserves, where observers were wardens and counts were completed along with other duties. Butterfly Conservation (BC) began a similar scheme in the early 1980's, where sites were self-selected by volunteers.

The aim of modelling the count data in the butterfly monitoring scheme is to

gain information on trends over years for different species of butterfly - to describe patterns in counts of butterflies and to estimate an index of butterfly abundance allowing for missing counts given a set of covariates - day of the year, wind speed, temperature, sunshine, grid reference, altitude, habitat-type and time of day. It is important to have information on abundance of any species, from a wildlife conservation point of view, in order to make informed management decisions and to identify species which require particular attention. This is especially the case for butterflies, as they are used by the British government as an "indicator" species of interest. Indicator species are chosen as they represent key species of interest within a habitat - they reflect the overall condition and quality of their ecosystem.

The different analysis methods used historically for estimating within and between year trends in animal abundance include linear interpolation (Pollard [1977]), log-linear models (Thomas and Martin [1996]), smooth models (Rothery and Roy [2001]), mixed models (Link and Sauer [1997b]) and generalized estimating equations (Link and Sauer [1997a]).

It was recognised that the linear interpolation method was not ideal, and an alternative method using smoothers was used by Rothery and Roy [2001]. They used Generalized Additive Models (GAMs) and modelled site-level trend as a smooth, nonlinear function. Model covariates could also be included. The flexibility of models such as GAMs, is important for data of these kind; butterfly species may show single, double or triple peaks in abundance, sometimes with rapid changes. GAMs also provide a framework for testing the statistical significance of changes in abundance provided important assumptions, such as independence of observations, hold. This assumption however, may be invalid and unrealistic for data such as these since counts collected within transects over time are likely to be temporally correlated and this correlation is unlikely to be explained, in full, by model covariates.

To address the requirements for both flexibility and potential dependence in the data, a model which allows for these important features was used for the analy-

ses that follow. Regression splines are used to provide this flexibility (similar to those used for GAMs) and Generalized Estimating Equations (GEEs) are used for fitting the model to accommodate any dependence in the data.

In this thesis, the issues of model selection, dependence within sites, detectibility and flexibility are addressed. I will also discuss methods for estimating annual indices of abundance at site and regional level, methods for calculating variance around these estimates, and methods for comparing these indices over time. Results are presented for some real data collected for the Butterfly Monitoring Scheme.

This thesis is structured as follows: Chapter 2 contains information on wildlife surveys undertaken in the past - mainly butterfly and bird surveys. It describes field methods and statistical methods used to analyse data collected. Chapter 3 describes the statistical theory behind the models used for analysing the BMS data, including linear models, GLMs, GAMs, mixed models and GEEs. Chapter 4 describes the particular statistical methods used to analyse our BMS data, and Chapter 5 presents results from having applied these methods to a subset of the BMS data.

Chapter 2

Review of Survey Methods for Butterflies and Birds

Longitudinal data of these kind (observed counts made on different sites i within a region over time t) have been collected many times before for many different purposes. The aim of this chapter is to look at different surveys of this kind which have been undertaken in the past, and to look at the different analysis methods used, usually dependent on the question of interest.
The main issues to be dealt with are:

- To calculate reliable annual (relative) indices of abundance at site level,
- To collate these site-level indices to produce regional and national indices,
- To compare these indices over time at site, and regional level, and to try to make inferences about the wider countryside.

Issues to keep in mind when trying to model the above are:

- Standard errors on estimates of index or trend
- Inclusion of covariates
- Observer effects

- Overdispersion
- Correlation
- Linear/smooth models
- Detection
- Model Selection
- Treatment of missing observations
- Requirements for inclusion of sites.

These items are addressed in the following sections with regards to surveys and analyses previously carried out, mainly in the United Kingdom and the U.S.A.. The main methods used historically to describe trends across years are the chain or ratio method, and the log-linear site by year effects model.

Data of these kind are known as longitudinal or repeated measures data (Chapter 3, Section 3.2), where measurements (or counts) are made repeatedly on sites (or in medical terms, subjects). Methods used need to account for the non-independent nature of the data. In the case of BMS data, counts are made on site i at time (Butterfly Monitoring Day) t. Over time, methods have been developed to model the time-series nature of this data and calculate indices of abundance. Some of these methods are addressed in this review.

Although this thesis deals with butterfly data, historically, there is only limited information available on analyses of butterfly data. For this reason, bird surveys are considered here also, as these often resemble those of butterflies. The main difference in field methods is that birds can be detected aurally as well as visually. However, analysis methods can be very similar.

2.1 Calculating within year Indices

2.1.1 Birds

The simplest case of calculating an index of abundance is that of the North American Breeding Bird Survey (NABBS), where only one survey is undertaken annually on each route (site). This scheme began in 1966 in order to develop relative indices of abundance of all breeding birds across the continental USA (now including parts of Canada and Mexico), and to compare these indices across space and time. Non-random roadside routes (or sites) were selected within geographical regions, with each route being 24.5 miles long. Stops are made every 0.5 mile, and point counts carried out for 3 minutes, with all birds detected aurally and visually within 0.25 mile recorded. Each route is visited at least once (but usually only once) during the breeding season (usually June, although May in some southern states), and the counts from the 50 stops along the route are summed to give an annual "index" of abundance. If more than one visit is made, the arithmetic mean of these is taken. Counts begin one half hour before sunrise, and usually last about 6 hours. As survey conditions are made as homogeneous as possible, no covariates are included in any of the models. It is assumed that once environmental covariates, e.g., temperature and sunshine level are at a certain level, counts will be independent of them. This assumption is discussed in Mountford [1982].

The United Kingdom Common Bird Census (UKCBC) was developed at the same time as the NABBS. It began in response to the global concern over the effect of increased usage of pesticides in the countryside on the population of birds throughout Britain. The B.T.O. (British Trust for Ornithology) needed a method of monitoring changes in population size of common bird species. A number of plots were established in 1962, with considerable plot turnover. More than 1500 plots have been surveyed, though not all continuously. The sample plots included covered a wide range of habitats, though mostly they were farmland or woodland sites, and predominantly in the South-East (Upton [1981]),

which makes it difficult to draw inferences beyond the surveyed habitat types to the wider country-side. Frequent visits (usually 6) were made by trained volunteers to self-selected plots and records were made of all birds of all species detected within the plot. Estimates were made of the number and location of territory holding males of each species. The method of estimation is described in Marchant et al. [1990] and an index is calculated by mapping successive visits during the breeding season (March-July). The maps were analysed by people to assess the number of territories present. The method is very time consuming and as the number of plots increased, a more automated method was required. As always, there are many methods to estimate the annual index - one example of a cluster analysis method of estimation is found in North [1977], although this method was found to be limited to only some species. As always, an objective, automated method is desired, with some level of estimation of precision.

In response to legitimate concerns raised with regard to the limitations of the CBC, in the 1990's, the BTO developed a new Breeding Bird Survey (UKBBS) in order to monitor changing population sizes (Newson et al. [2005]). BBS employs a formal, stratified, random sampling scheme of 1km squares in the United Kingdom. All adult birds detected (visually and aurally) are recorded (not just territory holding males). Each bird is assigned to one of three distances - 0-25m, 25-100m and $\geq$100m, or as in flight. Generally, two counts are made by skilled volunteers on each site annually and indices are calculated by averaging these visits. Habitat and species specific detection curves are modelled. Since sites are chosen randomly, any changes noted on sites can be extended to the wider countryside (or to that particular habitat). It can produce national indices by averaging regional densities, stratified by observer density. This survey scheme is the only one detailed in this review which attempts to address the issue of detectability (Chapter 3, Section 3.17).

2.1.2 Butterflies

The main method used by the BMS to survey butterflies in Britain is known as the Line Transect method or Pollard-Yates walk, and is described in Rothery and Roy [2001]. These terms are often used interchangeably, however, it should be noted that in Statistics literature, the Line Transect method implies randomisation of transects and distances of detection from the line were recorded, which is not the case in this survey. Due to the high variation in butterfly voltinism and phenology, more than one or two visits per season need to be carried out in order to obtain a reliable estimate of abundance, unlike breeding bird surveys such as those described in Section 2.1.1. Counts are made using the Pollard-Yates method, established by E. Pollard in the early 1970s at Monks Wood. The national scheme began in 1976. The method is based on transect counts, where the observer walks slowly and carefully over a pre-selected path of fixed length and width (usually 5m, or the natural boundary of the path). Any butterflies seen are recorded, and at the end of the walk, the number of each species seen is summed. The walks, or transects are governed by the following rules to provide standardisation and comparability between sites: Recording is carried out during the 26 weeks from the start of April until the end of September (although in recent years in the warmer southern sites, records have been made during March and October).

1. Counts are started after 10.45 am British Summer Time and completed before 15.45 pm.

2. Counts are not made when the temperature is below 13C; from 13C to 17C, counts are made in sunny conditions (60% sunshine minimum); above 17C, conditions for counts may be sunny or cloudy. On northern and western upland sites the minimum temperature in sunny conditions is 11C.

Every week, the mean counts per transect are calculated, and if only one count is made per week (as recommended) then that count is taken as the weekly site-level count. The counts for each week are summed for each site for each

year, giving an annual, site-level index, I_i of abundance for each species.

$$I_i = \Sigma_{t=1}^{26} y_{it} \tag{2.1}$$

where y_{it} is the count on site i at week t ($t = 1, ..., 26$). This is not a measure of population size, rather just an index of relative abundance which can be used to estimate trends over years. Unfortunately, the rules as laid out above are not always followed - e.g. weather patterns may disallow some counts, especially in Northern sites where ideal weather conditions are not always found, especially during March and/or October. Also, as with most volunteer schemes, missing counts are found due to volunteers taking holidays, or due to sickness, without any cover.

In the past, these missing counts have been estimated by interpolating linearly from the counts for dates either side, and finding the mean for the missing weeks.

$$y_t = \frac{1}{2}(y_{t-1} + y_{t+1}) \tag{2.2}$$

where y_t is the site-level count for week t. This method often misses peak counts which also means that indices of abundance may be biased low. This is especially the case when more than one week has been missed consecutively over a season, which unfortunately and unavoidably happens fairly regularly.

It was recognised that the linear interpolation method was not ideal, and an alternative method using smoothers was used by Rothery and Roy [2001]. They made use of GAMs (Chapter 3, Section 3.10) to predict the missing counts. Each site has its own log-linear regression model where the expected count in week t is

$$E[Y_t] = \mu_t = \exp[f(t; d)] \tag{2.3}$$

where f(t;d) is a function denoting a cubic smoothing spline with d degrees of freedom (to be selected usually using the data-driven method of cross-validation). The flexibility of models, such as GAMs, is important for data like these; but-

terfly species may show single, double or triple peaks in abundance, sometimes with rapid changes. GAMs also provide a framework for testing the statistical significance of changes in abundance provided important assumptions, such as independence, hold. This assumption however, may be unrealistic for data such as these since counts collected within transects over time are likely to be correlated and this correlation is unlikely to be explained, in full, by model covariates.

Brown and Boyce [1998] describe a method for estimating abundance and density of the Karner blue butterflies in Wisconsin, USA. They make use of Distance Sampling (Buckland et al. [2001]), a method widely used to survey animal populations. Multiple surveys are undertaken at sites throughout the target area (across Wisconsin), and detection curves (Chapter 3, Section 3.17) estimated. This method provides an asymptotically unbiased estimate of true abundance, allowing for covariates, once certain assumptions hold.

To address the requirements for both flexibility and potential dependence in the data, methods which allow these features were used for this analysis. In the analyses that follow, splines are used to provide this flexibility (similar to those used for GAMs) and Generalized Estimating Equations (GEEs) are used to accommodate any dependence in the data.

2.2 Collating Indices

Once a site level annual index has been calculated for a survey, the next step is to collate these indices (by region or habitat type) and compare them over time. A method is required which produces unbiased estimates and which corrects for effort.

2.2.1 Birds

The main method used to collate site level indices to regional ones is just to sum the site level indices using some kind of weighting scheme. In the NABBS, site

level counts are "averaged" to produce an annual index. Collation is usually weighted by the proportion of site area, though it also could be weighted by abundance or precision. In the UKCBC, site level indices (or number of territories) are simply added up to produce regional indices. No account is taken of site area. For the UKBBS, detectability is measured at a regional (habitat specific) level to produce national population and density estimates.

2.2.2 Butterflies

Once relative indices of abundance are calculated, they need to be collated over sites. The most simple method of doing this is just to add the different site indices to produce a regional index. This method though, is limited, as the regional index can be swamped by one particular site having a particularly high count due to natural variation. This problem is dealt with in Moss and Pollard [1993]. It suggests logging the site indices before adding them. This method has the advantage of making it less likely to see significant changes over years when changes are only due to variation and not due to spurious high counts at one particular site. High counts are often due to site area, and this should be dealt with by using an offset of transect area in the model.

2.3 Comparing Indices over time

The main aim of population monitoring is to model trend - that is - to know if a population is increasing or decreasing, at site, regional and national levels. Managers need to know if populations are in danger of extinction. There are many different ways to model this, often producing different results. These are discussed in the following sections.

2.3.1 Birds

Broadly, there are two modelling philosophies in comparing indices over time. The first of these is the chain, or ratio method. Mountford [1982] describes this

method of comparing counts across years. Let I_{i1} and I_{i2} be site-level indices on sites i in years 1 and 2.

$$r_{21} = \frac{\Sigma_{i=1}^{K} I_{i2}}{\Sigma_{i=1}^{K} I_{i1}}, \tag{2.4}$$

where the ratio is defined as the ratio of the sum for each year, over the K sites surveyed in both years. To look at trends over longer periods of time, say four years - $r_{41} = r_{43}r_{32}r_{21}$, which for balanced data is equal to $\frac{\Sigma I_{i4}}{\Sigma I_{i1}}$. Unfortunately, as for most surveys of this nature, the same sites are not included every year, so the data are unbalanced. This needs to be accounted for. Also, the data are highly serially correlated, with animals displaying high levels of site fidelity (Mountford [1982]). This method uses no covariates - environmental conditions are assumed to be constant over time. This is a crucial assumption and can be tested. In one example, 9 years is taken as the limit of stable conditions.

The chain method is rather limited, as it makes no attempt to fill in missing data. It is an inefficient method, and much work has been done in attempts to improve on it.

The main method used for trend estimation in the NABBS is linear route regression (Thomas and Martin [1996]). Route (or site) level trends are calculated by logging the index I, and calculating linear trends over time (years).

$$log(I_t + c) = \beta_0 + \beta_1 t + \varepsilon_t \tag{2.5}$$

where I_t is the index for a site in year t and c is some constant. β_1, the slope of the line, is taken as the site level trend. These site level trends are then averaged to produce regional trends. Details of this method differ between the American and Canadian analyses, as different organisations administer the surveys - the US National Biological Service in America, and the Canadian Wildlife Service respectively. For example, the transformation of indices is either log(index+0.5) or log(index+0.23). Back transformations can be performed on either the site level trend, or on the regional trend. Averaging can be weighted by either precision, area or by abundance. Covariates such as observer effects can be

included. The differences in details can have important consequences in the statistical significance of trends for different species. This method also depends on the assumption that the relationship between year and logged count is linear - it does not allow for any other shape.

A similar method to this is the use of a site by year effects model using Poisson regression.

$$log(I_{it}) = \alpha_i + \beta_t + \varepsilon_{it} \tag{2.6}$$

where α_i is the site effect for site i and β_t is the year effect for year t. This method allows for a site and a year effect, so allows for fluctuations other than linearity on the log scale. It uses a separate parameter for each site and year effect, and missing values can be imputed easily. It is dependent on the question of interest whether a long range "average" (i.e., linear) trend or a completely unconstrained method is preferred. This question is addressed in Fewster et al. [2000]. GAMs (see Chapter 3, Section 3.10) are a useful tool in summarising this. Here, the level of smoothing can be chosen by the analyst - usually a balance between linearity (on the log-scale) and capturing every single fluctuation is desired. The model used in Fewster et al. [2000] is

$$log(\mu_{it}) = \alpha_i + f(t) \tag{2.7}$$

where μ_{it} is the expected index in site i in year t and $f(t)$ represents some smooth function of time estimated from the data. The level of smoothing depends on the degrees of freedom used - a degree of freedom of 1 corresponds to a straight line, as in the NABBS analysis, and a degree of freedom for every year surveyed corresponds to the site-by-years model as in Equation 2.6.

All of these models assume that counts are temporally and spatially independent. Use of these models infers that changes across years are measured by

$$r_t = \frac{I_t}{I_1} = \frac{exp(\beta_t)}{exp(\beta_1)} \tag{2.8}$$

or

$$r_t = \frac{I_t}{I_1} = \frac{\exp(f(t))}{\exp(f(1))} \tag{2.9}$$

where I_t is the total (predicted) for year t, $f(t)$ is the smooth function of time, and the base year is taken as year 1, although in practice, any year can be taken as the base year. One important aspect of using the GAM method is to decide on a smoothing parameter. The objective is to capture all major features of population trend, while ignoring yearly fluctuations. Precision is calculated using bootstrapped sites. Covariates are easily implemented into the GAM framework. Similar results to these are encountered when polynomial Poisson models are used. If sites are randomly selected, GAMMs could also be investigated. Observer effects are discussed in Link and Sauer [1997b]. Many studies have shown that between and within observer effects can have significant effects on the conclusions of an analysis. These can easily be incorporated into the GAM, GLM, GEE or GAMM framework. Another method used to describe trends but not the magnitude of the trend for the NABBS is non-parametric rank-trend analysis (Thomas and Martin [1996]). Counts for each route are arranged in ascending order and assigned ranks R_i. A test statistic, D, is calculated for each route as $D = \sum_1^t (R_i - i)^2$, for t years. If D is small, counts are increasing, and if D is large, counts will tend to decrease. This is a simple and useful tool to test for the significance of trends, though it only considers those increasing or decreasing and does not allow for other fluctuations.

2.3.2 Butterflies

The butterfly monitoring scheme in the Netherlands (DBMS) has been running since 1990 and surveys between 100 and 300 sites in any one year. De Vlinderstichting uses similar surveying methods to BMS (van Strien et al. [1997]), in association with C.B.S. (Statistics Netherlands). Weekly counts are made, and the sum of these weekly counts (area under the curve), over the entire surveying season, is taken as the annual index for that species for that site. Again,

sites are self-selected by volunteers, making wider inference problematic. As with the BMS, the aim of De Vlinderstichting is to monitor long term trends in butterfly populations, and to use annual indices to enable local management decisions. Changes in abundances are measured as indices using the software TRIM (Trends and Indices for Monitoring data). Indices are then stratified as in van Swaay et al. [2002].

TRIM (Pannekoek and van Strien [2005]) is used as a tool to analyse wildlife monitoring data. It allows for missing counts and uses loglinear regression, as in Equation 2.6, to estimate site-level annual indices. It presents results as annual trends, allowing for the effects of covariates. It also allows for overdispersion, which is generally important in Poisson count data. Observations are inputted into TRIM as I_{it}, an index for site i in year t. Trends, or indices relative to other years for site i in year t are calculated as $\frac{I_{it}}{I_{i1}}$ - all being compared to year 1, as the first year is taken as the baseline year. Different classes of models are permitted to calculate the trend - ranging from a very simple sites-only model, with $log(\mu_{it}) = \alpha_i$ for all years t, to a more complex model, allowing a slope time effect, with L change points, $l = 1, ..., L$:

$$log(\mu_{it}) = \alpha_i + \sum_{l=1}^{L} (\beta_l - \beta_{l-1})(t - k_l)D(t, k_l) \tag{2.10}$$

where $D(t, k_l) =$

- 0 for $t \leq k_l$
- 1 for $t > k_l$,

and $\beta_0 = 0$, for time-points k_l and change-points l.

These models assume a common year effect for each site, but site-specific categorical covariates can also be used in the model (e.g. habitat type). Trend over years can be explained by an ordinary least squares estimator of the slope parameter of a linear regression line through the logged yearly totals. If this is significantly different from 0, a change over time is thought to be significant.

Another method commonly used to analyse butterfly data is the site by year model, as in Equation 2.6.

2.4 Discussion

The methods above all depend on satisfying certain assumptions, with varying consequences. For example, for the NABBS analyses, the trend is assumed to be linear on the log scale. This is not always satisfied, especially in the case of ecological data, which can be very unpredictable and non-linear. For this reason, GAMs are investigated. Independence is assumed for most of the models discussed so far. Again, in reality, there is correlation within sites (and sometimes across sites). The impact of this correlation on estimated indices and trends should also be investigated.

So far, I have discussed obtaining within year indices for the species of interest, collating these indices, and comparing these indices across years. I have not yet investigated the precision of these estimates. Model comparison and selection are other important topics demanding attention. Widely used model selection methods such as Pearson's chi-squared statistic, AIC and the likelihood ratio test all assume independence. If this is not satisfied, it could have important consequences for the covariates included. This, in turn could lead to quite different relative indices being estimated. Methods of model selection which relax this assumption should be considered.

Chapter 3

Statistical Theory Underlying the Models Introduced

In this chapter, I will introduce and describe the notation and theory regarding the statistical models and methods used in this thesis to analyse the butterfly survey data. The models include Generalized Linear Models (GLMs), Mixed Models, Generalized Additive Models (GAMs), regression splines and Generalized Estimating Equations (GEEs).
The chapter is divided into:

- Describing the class of data collected for this thesis,
- Describing the models used to investigate the data,
- Describing the methods of estimation used within the models.

As always, the method chosen to model our BMS data is dependent on the question of interest we want to answer. We are primarily interested in developing a regional model, averaging over the population of interest, and predicting daily counts at individual sites. Correlation within sites is common for this kind of

data set, so this must be catered for, although we are not strictly interested in the form of the dependence. As with all ecological count data, our data are highly overdispersed and non-linear, so these issues also need to be accounted for. Different models have different properties - both advantages and disadvantages, and rely on different model assumptions. This chapter aims to describe the models under consideration and list the relative worth of each.

3.1 Notation

The notation used follows the standard notation used in the literature.
Lower-case letters describe observed values $y_1, .., y_n$, which are realizations of the random variables $Y_1, ..., Y_n$. Greek letters denote parameters, with a hat-symbol used to denote parameter estimates. Bold, underlined text indicates vector and matrix terms.
The following lists the notation used in this thesis:

- y_{it} is the observed butterfly count on site i at time (BMSDay) t
- β_0 is an intercept term
- $\underline{\beta}$ is the vector of model parameters
- $\underline{\hat{\beta}}$ is the vector of the estimated model parameters
- t is the time of year (i.e. Butterfly Monitoring Day) - which runs from Day 1 to Day 250 - March 1st to October 31st, though sometimes, t refers to year or week t.
- μ_i is the mean of site i
- variance is represented by σ^2
- X_{it} represents the matrix of covariates
- ε_{it} represents model errors
- g(.) is the link function

- $V(\mu)$ is the variance function
- ϕ is the dispersion parameter
- K represents the number of sites within a region
- p is the number of regression coefficients to be estimated, also usually equal to the model degrees of freedom
- R is the working correlation matrix
- $Var(Y_{it}) = V(Y_{it}) = V(Y_i) = V_i$ is the variance of the observations, the $n \times n$ variance-covariance matrix
- A_i is the diagonal matrix with elements $Var(Y_{it})$
- $V(\hat{\beta})$ is the variance of the parameter estimates
- L is the model likelihood
- l is the model log-likelihood.

3.2 Longitudinal Data

Diggle et al. [2002] introduce the topic of Longitudinal data. Traditional data analysis depends on the assumption of independence. Data are collected and analysed, and inferences are made. However, models in medicine and ecology often do not satisfy this key independence assumption.

In pharmaceutical or medical trials, subjects are conventionally assigned to a treatment, and followed over time. The within-subject effects must be somehow separated from the between-subject effects. The temporal nature of the data must also be addressed in any analysis. Time series is an area of statistics for which standard methods have been developed, but time series data usually are data collected on one variable, over a long period of time (several years), whereas longitudinal data have many subjects, followed over relatively short periods of time (often within one year). Therefore, longitudinal data must

address issues of temporal correlation, along with small numbers of independent (usually) subjects. Methods developed for dealing with these type of data are borrowed from traditional data analysis of independent data, and time series methods.

The following sections describe the models developed to analyse longitudinal data, along with estimation methods.

3.3 Linear Models

I begin with reviewing the simplest model available - the simple linear model. For simple linear regression,

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t \tag{3.1}$$

Here, the Y_t's are the response - the observed count made on Butterfly Monitoring Day t, β_0 is the intercept parameter (expected count when $t = 0$) and β_1 is the slope parameter (expected change in count with unit increase in time), to be estimated, usually using the method of Maximum Likelihood.

Another form this can take, allowing for a vector of covariates X is as follows:

$$Y_t = X\underline{\beta} + \varepsilon_t \tag{3.2}$$

For the linear model, the data are considered to be Normal, $Y_t \sim N(\mu, \sigma^2)$, i.e., the observations Y_t are normally distributed with mean μ and variance σ^2.

The assumptions made here are that;

- there exists a linear relationship between the Y_t's and the covariates X,
- there is a constant variance, σ^2,
- the Y_t's are independent (i.e., uncorrelated),
- the errors are normally distributed, i.e., $\varepsilon_t \sim N(0, \sigma^2)$

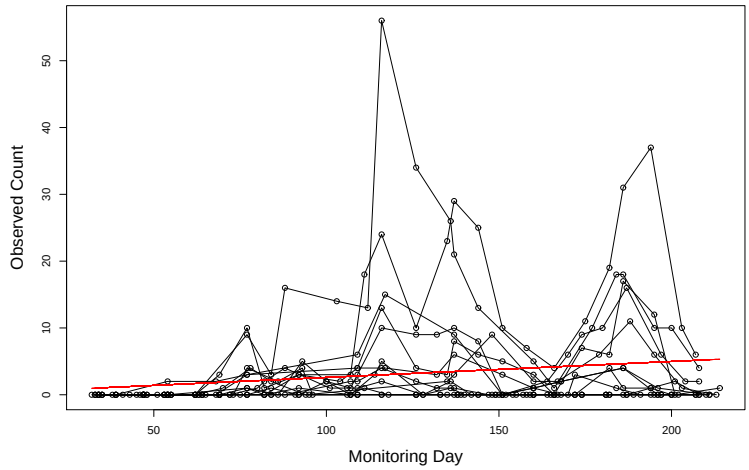


Figure 3.1: Weekly Observed Counts for South Eastern Region, 2002, with best-fit straight line through (in red). Monitoring Day runs from March 1st to October 31st.

A simple linear model can also be developed at regional level, where several sites are modelled at once:

$$Y_{it} = \beta_i + \beta_1 t + \varepsilon_{it} \tag{3.3}$$

This model assumes a regional level slope parameter, but allows a separate intercept parameter for each site.

Clearly, as seen in Figure 3.1, this method would be entirely useless to produce a realistic index of abundance. Figure 3.1 shows the best fit straight line through a plot of observed count against day for data from the South East region, to be described more fully in Chapter 5, Section 5.1. Often, however, as in the case of the BMS data, normality cannot be assumed, and we need to restrain our response. To do this, an extension of these linear models - Generalized Linear Models, was developed.

3.4 Generalalized Linear Models

GLMs were developed as an extension to linear models, to allow for more complex relationships between the response and the explanatory variables, e.g. binary or count data.

Generalized Linear Models have three main components:

- a family, or distribution (the exponential family, Section 3.4.1 includes all the standard distributions used in GLMs),
- a linear predictor,
- a link function.

Instead of having

$$E(Y_{it}) = \mu_i = X_{it}^T \underline{\beta}, \tag{3.4}$$

we now have

$$E(Y_{it}) = \mu_i \tag{3.5}$$

and

$$g(\mu_i) = \eta_i = X_{it}^T \underline{\beta}$$

where *g(.)* is a monotone link function.

The main assumptions involved with GLMs are as follows (Hardin and Hilbe [2001]):

- that the Y_{it}'s are independent (i.e., uncorrelated),
- that the variance function $V(\mu)$ is correctly specified,
- that the dispersion parameter ϕ is correctly specified (i.e., is equal to one for Binomial and Poisson data), and,
- that the link function is correctly specified.

Linear models are a special case, when the link function is the identity link and the distribution is normal.

For the Poisson distribution, the natural, or canonical link function is the log link. This is chosen as the range of values that can be taken lie from zero to ∞, i.e., values must be non-negative integers. GLMs do not assume constant variance, but assume that there is a known relationship between the mean and variance. They also assume linearity on the scale of the link function. GLMs solve the problem of non-normality and non-constant variance, but there are still issues of non-independence in our data to deal with, and whether linearity on the scale of the link function is a reasonable assumption. GLMs, however, are still a good framework to begin with for analysing the BMS data.

3.4.1 Exponential Families

Most of the useful one parameter distributions used in modelling belong to what is called the general exponential family.
The random variable Y has a distribution belonging to a one parameter exponential family if it has a density or probability function

$$f_y(y;\beta,\phi) = \exp\left\{\frac{y\theta - b(\beta)}{a(\phi)} + c(y,\phi)\right\} \tag{3.6}$$

or equivalently, its log-likelihood is:

$$l(\beta) = logL(\underline{\beta},\phi;y) = log(f_Y(y;\underline{\beta},\phi)) = \sum_{i=1}^{n}\left\{\frac{y_i\beta_i - b(\beta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{3.7}$$

where:

- n is the number of observations,
- β is the natural or canonical parameter,
- ϕ is the scale or dispersion parameter, and
- a, b and c are known functions.

Two important functions of the log-likelihood are the score function (Equation 3.8) and the Fisher information (Equation 3.10).

The score function U is the partial derivative of the log-likelihood function of Y with respect to the parameters, and is denoted by $U(\beta) = \frac{\delta}{\delta\beta} l$

An important result of this is as follows:

$$E(U) = E_Y\left(\frac{\delta}{\delta\beta} l\right) = 0 \tag{3.8}$$

Also, if we partially differentiate this again with respect to β, we get

$$V(U) = E(U^2) = E(-U') \tag{3.9}$$

or

$$E_Y\left(\frac{\delta^2}{\delta\beta^2} l\right) + V\left(\frac{\delta}{\delta\beta} l\right) = 0 \tag{3.10}$$

or $E(-U') = E(U^2)$. This gives us that the Fisher information $I(\beta) = E\left[(\frac{\delta}{\delta\beta} l)^2\right] = -E(\frac{\delta^2}{\delta\beta^2} l)$

Therefore,

$$E(Y) = \mu = b'(\beta)$$

and

$$V(Y) = a(\phi)b''(\beta). \tag{3.11}$$

This leads to $V(\mu) = \frac{V(Y)}{a(\phi)}$ where $V(\mu)$ is the variance function and $a(\phi)$ is the term to allow for dispersion - taken as one for the Binomial and Poisson families. We can use this exponential family to show how generalized linear models are formed for the Poisson distribution in the following section.

Maximum likelihood parameter estimates are found by solving the Score equation

$$\frac{\delta}{\delta\underline{\beta}} l = 0, \tag{3.12}$$

or in matrix notation, by solving the set of score equations;

$$U(\underline{\beta}) = \frac{\partial l}{\partial \underline{\beta}} = \sum D_i^T V_i^{-1}(y_i - \mu_i) = 0, \tag{3.13}$$

where D_i is the matrix of derivatives with elements

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_k} x_{it}$$

and V_i is diagonal with elements $V(Y)$. This is usually not solvable by analytic methods, and needs to be estimated using either Newton-Raphson method or the Iterative Re-weighted Least Squares (IRLS) method. The variance-covariance matrix of the parameter estimates is usually based (analytically or numerically) on the Hessian matrix - it is important to note, though, that the variance matrix is estimated after the calculation of $\underline{\hat{\beta}}$, the parameter estimates.

Poisson Distribution

As the data are in the form of counts Y_{it} on sites i at times (days) t, the Poisson GLM is the most suitable (though still assuming that counts made on a site are independent). The probability distribution is as follows:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}, \tag{3.14}$$

for $y \geq 0$ or written in a form to make it comparable with Equation 3.7, the log-likelihood becomes

$$log(l) = \frac{ylog(\mu) - \mu}{1} - log(y!) \tag{3.15}$$

The denominator of 1 refers to the dispersion parameter $a(\phi)$ of Equation 3.7. For the Poisson distribution, we are allowed to have some Y_{it}'s equal to zero, however, the expected value $E(Y_{it}) = exp(X_{it}\beta)$ must be greater than zero. So, we have that the natural, canonical parameter (the link function) is $log(\mu)$, and

the dispersion parameter $\phi = 1$. This means that we are assuming that the variance equals the mean, for observations (counts) y_{it} for monitoring days 1 to 250, and all surveyed sites K within a region, a fact which is rarely true with real-life Poisson data. For example, for a section of the BMS data, the mean of the counts for the South East region is 3, whereas its variance is 44. In reality, the variance is much larger than the mean, i.e., $V(Y) > E(Y)$.

In more formal notation, the full log-likelihood for the Poisson distribution is

$$L(\beta|X, y_i, ..., y_n) = \sum_{i=1}^{n} \{-\exp(x_i\beta) + y_i x_i \beta - \ln \Gamma(y_i + 1)\},$$

where $g(\mu_i) = ln(\mu_i) = X_i\beta$. In solving for β, we need to solve (as in Equation 3.12), which becomes,

$$\left[\left\{ \frac{\partial L}{\partial \beta_t} = \sum_{i=1}^{n} (\frac{y_i}{\mu_i} - 1)(\frac{\partial \mu}{\partial \eta})_i x_{ti} \right\}_{t=1,...,p} \right]_{p \times 1} = [0]_{p \times 1}$$

where $(p \times 1)$ refers to estimating the $(p \times 1)$ vector of the p β parameters.

However, these parameter estimates are still built upon the assumptions of independence between counts, and linearity on the scale of the link function. A more sophisticated analysis method should be used to account for these issues.

3.5 Overdispersion

One of the earliest methods for dealing with variance greater than expected in Poisson (or Binomial) data was to calculate a dispersion parameter ϕ. This allows for the usual GLM estimation of parameters, and afterwards, adjusts the standard errors of the parameters by multiplying them by ϕ. ϕ is usually calculated by dividing the deviance statistic by the residual degrees of freedom (n-p). This is an *ad hoc* method of dealing with the problem. Overdispersion is caused in cases like those of the BMS, where the variance exceeds the mean, however, underdispersion is also possible, e.g. in the case of animal litters, the variance within a litter may be smaller than predicted in a model.

GEEs (see Section 3.8) offer a more unified method, which accounts for the simultaneous estimation of correlation and model parameters.

3.6 Quasi-likelihood

All of the above GLM theory depends on choosing a distributional form for the data (e.g., Binomial, Gaussian or Poisson) and deriving a likelihood function with its resulting theoretical properties. Often, though, the observed data do not correspond to any distribution exactly, and so we cannot rely on the maximum-likelihood function for estimation. For this reason, an extension was developed - the quasi-likelihood function, where only the relationship between the mean and the variance of the observations needs to be specified.

Define the quasi-likelihood function $Q(y_i, \mu_i)$ as:

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \tag{3.16}$$

or equivalently

$$Q(y_i, \mu_{i'}) = \int^{\mu'_i} \frac{y_i - \mu'_i}{V(\mu'_i)} d\mu'_i + f(y_i).$$

Wedderburn [1974] describes how estimation using maximum quasi-likelihood is directly equivalent to estimating using maximum likelihood, without having to rely on choosing the correct distribution for the observed data. Nelder [2000] acknowledges that one of the most important quasi-likelihoods is that for the overdispersed Poisson distribution. If the link and variance correspond to a particular member of the exponential family, then the quasi-likelihood is equal to the likelihood proper. The issue of quasi-likelihood becomes important in Section 3.12 on model selection.

For the Poisson distribution, the quasi-likelihood is:

$$\sum_{i=1}^{n} \{y \ln \mu_i - \mu_i\}. \tag{3.17}$$

for n observations.

3.7 Pseudo-likelihood

Pseudo-likelihood is another estimation method discussed in Nelder [2000] and Wolfinger and O'Connell [1993]. It is often confused with quasi-likelihood (Section 3.6). Pseudo-likelihood is related to quasi-likelihood, in that it allows for a relationship between the mean and the variance; however, pseudo-likelihood bases its estimation on the Normal distribution. It allows the variance function σ^2 to be a function of the mean μ. Pseudo-likelihood assumes normal errors, though allows the variance to change with the mean. This assumes that the errors are symmetric, though this is rare in practice. It does not differentiate between the variance function $V(\mu)$ and the dispersion parameter ϕ, as is the case of GLMs.

3.8 Generalized Estimating Equations

As stated previously, we cannot rely on using GLMs, as some of the assumptions are not valid. Another method is required, which will allow for correlation in the data.

Longitudinal data of this kind, which follows a site (or patient) over time (i.e., repeated measures), are very common, especially in biomedical sciences, and new methods are being developed to deal with them. Section 3.2 expands on the topic of longitudinal data. In order to develop an index of abundance for site i at time t, a new set of models is necessary, since classical linear models and GLMs are unsatisfactory. GEEs are extensions of GLMs which allow for:

- Longitudinal type data, following a site over time,
- Correlation among observations on the same unit (site)

Hardin and Hilbe [2003] introduce the Generalized Estimating Equation (GEE) with regards to panel data, i.e., where data are clustered (e.g., litters of animals)

or repeated measures (e.g., a patient followed through time). GEEs are split into subject-specific (SS) models or population averaged (PA) models (also known as marginal models). The former models individual panels (sites) and attempts to explain the source of covariance, the latter's regression coefficients describe the average population response and only describe the covariance among repeated observations (do not attempt to explain it). Since we only want to allow for dependence and are not primarily interested in it scientifically, we will use a PA model. In the case of normal data, there is a simple relationship between SS and PA models, but for other distributional forms, this becomes more complicated. Although attempts were made to solve the problem of correlated data for several individual GLM models, GEE is a broad framework which unifies these methods.

One of the primary advantages of using GEEs is that no distributional form is assumed, so there is no danger of specifying a wrong distribution, and avoids the need to specify a multivariate distribution. GEEs are essentially a general, unified form of using quasi-likelihood in that there are no parametric assumptions.

As in the case of GLMs, we have the following scenario: We have a set of observations from K sites, and for each site i, $(i = 1, ..., K)$, we have a vector of observed counts y_{it}, for Butterfly Monitoring Survey days $t = 1, ..., 250$ with corresponding vector of covariates $X_{it} = (x'_{i1}, ..., x'_{i250})$, where some covariates are site-specific, and some change over time, for the 26 weeks in the Butterfly Monitoring Season, allowing for missing values. In general, we assume that components of y_{it} are correlated, but that individual sites are independent from one another - this issue is further discussed in Chapter 4, Section 4.3.7. To model the relationship between the count and the covariates, with the ultimate aim of deriving an annual, site-level index of abundance, a regression method similar to that of the GLM is formed:

$$g(\mu_{it}) = X_{it}\underline{\beta}$$

and

$$\mu_{it} = E(Y_{it}|X_{it}),$$

and

$$\underline{\beta} = (\beta_1, ..., \beta_p)$$

is a vector of p unknown regression coefficients to be estimated and $g(.)$ is the link function (usually taken as the canonical link function of a distribution).
The main difference between GLMs and GEEs is that the GEE allows R, a working correlation matrix, to be specified. Some general forms of this working correlation matrix are the independence structure, the exchangeable (or compound symmetry) form, or the auto-regressive form. There is much discussion in the literature about the consistency and efficiency of estimators of $\underline{\beta}$ under the different forms assumed, and the underlying truth of the correlation structure (e.g., McDonald [1993] and Fitzmaurice [1995]). Choosing the "correct" correlation structure, and model selection in general is discussed later in this chapter.
The form the variance of observations Y_{it} takes is as follows:

$$V(Y_i) = \phi A_i^{\frac{1}{2}} R A_i^{\frac{1}{2}}, \tag{3.18}$$

where ϕ is the dispersion parameter, and $A_i^{\frac{1}{2}}$ is diagonal matrix with elements $\sqrt{V(Y_{it})}$ and R is the working correlation matrix.
GEE theory is based on the assumption that there are no missing data, or if missing data exists, they must be missing completely at random (MCAR).
When fitting GEEs, the following items need to be considered:

- a model for the mean
- a model for the variance

It was Liang and Zeger [1986] who first introduced the unified GEE approach, and estimation is described in the following section.

3.8.1 GEE Estimation

Focus here is on PA-GEE estimation, which models the average count across all panels (sites), so that we develop a marginal regional model, which can predict counts at individual sites, using site-specific covariates. This is the brand of model introduced in Liang and Zeger [1986].

We want to estimate $\underline{\beta}$, the vector of parameter estimates and V, the associated variance-covariance matrix.

We want to solve equation 3.13, but instead of V being a diagonal matrix with diagonal elements $V(Y_{it})$, V becomes $V_i = (A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}})\phi$, where the A is the diagonal matrix with elements $V(Y_{it})$ and R is the correlation matrix (to be specified), and ϕ is the dispersion parameter.

Beginning with the assumption of independence between all observations, R becomes the identity matrix, and $V(Y_{it})$ is as before in Equation 3.11. However, we can specify R to have other forms, and the estimation algorithm estimates simultaneously the marginal model parameters as well as the correlation parameters. The method alternates between updating the estimates for $\underline{\beta}$ and updating the correlation matrix R.

3.8.2 Specifying the Correlation Matrix

An important aspect of the GEE is specifying the form of the correlation matrix, R. According to Liang and Zeger [1986], the GEE approach yields a consistent estimator of $\underline{\beta}$ even when R is misspecified. For this reason, an independence model is often used when the choice of R is not obvious. The most commonly used are:

- Independence model
- Compound symmetry (or exchangeable), where

$$corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{for } j = k \\ \rho & \text{for } j \neq k \end{cases}$$

- First order Auto-regressive (AR-1), where

$$corr(y_{ij}, y_{ik}) = \rho^{|j-k|}$$

- Unstructured - where every element of the correlation matrix is estimated separately.

Zeger et al. [1988] also demonstrated that $\hat{\underline{\beta}}$ obtained under the independence model is relatively efficient. This (independence) is a dangerous assumption to rely on, however, as McDonald [1993] showed that when the correlation between responses is large, $\hat{\underline{\beta}}$ becomes inefficient. Fitzmaurice [1995] discusses this further and concludes that for time-varying or cluster-specific covariates (both of which we have in the BMS data set), estimates of $\hat{\underline{\beta}}$ under the independence assumption may be very inefficient, often as low as 60%. Efficiency here is measured as Asymptotic Relative Efficiency (ARE), where the ARE of an element of $\hat{\underline{\beta}}$ is measured as the ratio of elements of the asymptotic covariance matrix of a model with correctly specified correlation and the the asymptotic covariance matrix given by the incorrectly specified "working" correlation matrix. Zeger et al. [1988] warns of the danger of ignoring correlation - in some studies, it leads to incorrect interpretations of data and model inferences. There is strong suggestion that there will be autocorrelation within site for the BMS data set, and it is important that this should be accommodated for, if found to be significant.

3.9 Mixed Models

Mixed models are an extension of simple linear models (Section 3.3) which allow for a more flexible specification of the correlation in the data. They are defined as:

$$Y_{it} = X\underline{\beta} + ZU + \varepsilon_{it}, \tag{3.19}$$

where $X\beta$ is as specified in linear models. Z is a design matrix and U is the vector of unknown random effects parameters. $\varepsilon_{it} \sim N(0, \sigma^2)$ as before, in

Equation 3.1. $X\beta$ is defined as the fixed effects parameters, with the random effects being ZU, a combination of both effects giving mixed effects models. $U \sim N(0, G)$ and, such that $E(U) = 0$ and $V(U) = ZGZ'$, giving $E(Y) = X\beta$ and $V(Y) = ZGZ' + \sigma^2$.

Mixed models are another method used to analyse longitudinal data, however, this basic mixed model assumes normality of the errors. Verbeke and Lesaffre [1997] demonstrate that maximum likelihood estimates of the parameters are consistent, even when the normality assumption is not valid, however, they are dealing with continuous longitudinal data and our BMS data are discrete.

Both Zeger et al. [1988] and Horton and Lipsitz [1999] observe that the linear mixed model corresponds to the SS model for the GEE with exchangeable working correlation and identity link - it explicitly models the random effects of the panels/clusters/sites with a parametric distribution.

Simple linear models can be considered special cases of generalized linear mixed models with $Z = 0$ and $R = \sigma^2 I$ and link function equal to the identity.

Multi-location trials in agriculture were the basis for the development of mixed models. Experimental trials at different sites correspond to our sites in the BMS data. The question of interest is whether the sites should be considered as fixed or random effects. Littell et al. [1996] discuss this issue. Locations can be considered fixed if they have been specifically selected for inclusion and have known characteristics. Inference can only be considered for the specific sites included. This means that we could not generalise any results to the wider country-side. Sites are considered random if they have been randomly selected from a wider population of sites. In our case, neither assumptions are valid.

Mixed models are a very useful tool in describing and analysing longitudinal data. However, the simple mixed models assume normality of the raw data, and we cannot identify our sites as either fixed or random. For these reasons, I would be very reluctant to use these methods to analyse our BMS data. However, GLMMs (Section 3.9.1) are an extension of mixed models which allow for non-normal data.

3.9.1 Generalized Linear Mixed Models

Generalized Linear Mixed Models are an extension of Mixed Models in an analogous way that GLMs are extensions of simple linear regression. GLMMs allow fixed and random effects to be estimated, whilst assuming that the data come from a distribution which is a member of the exponential family. The random effects allow the natural heterogeneity between panels (sites) to be modelled. The GLMM is specified as:

$$E(Y_{it}) = \mu_i \tag{3.20}$$

where

$$g(\mu_i) = \eta_i = X\underline{\beta} + ZU \tag{3.21}$$

and $\underline{\beta}$ are the fixed effects parameters to be estimated, **U** are those corresponding to the random effects. Estimation is either via conditional likelihood or marginal likelihood. Often, though, the estimation is numerically intractable (Breslow and Clayton [1993]). Estimation is often based on Pseudo-likelihood (PL) (Section 3.7), either PL or restricted pseudo-likelihood (REPL), where the difference is the same as that described in the next section.

Heagerty and Kurland [2001] warns of the dangers of using the GLMM when there are autoregressive random effects. This can lead to substantially biased point estimates of the fixed effects parameters.

3.9.2 Maximum likelihood versus Restricted maximum likelihood

Estimation in mixed models is either via maximum likelihood (ML), or restricted maximum likelihood (REML) - the latter taking into account the degrees of freedom needed to estimate the fixed effects in order to calculate the variance. REML estimates the random variance components unbiasedly whilst ignoring the fixed effects components.

For these reasons, parameters are estimated using ML, and REML are used

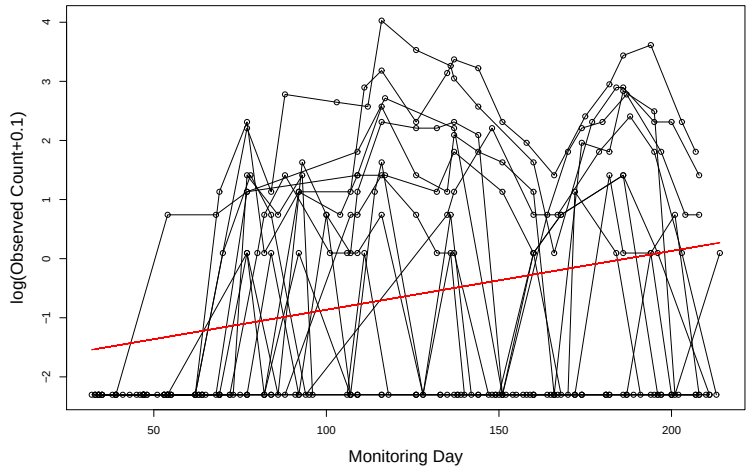


Figure 3.2: Weekly Observed Counts for South Eastern Region 2002, on the log scale, with best-fit straight line through (in red). Monitoring Day runs from March 1st to October 31st.

to estimate the variance components for the GLMMs considered in Chapter 5, Section 5.3.2.

3.10 Generalized Additive Models

All of the above models still assume that the response (count) is linearly related to the covariates on either the scale of the response, or on the scale of the link function. The relationship between count and day is clearly more complex than this (see Figure 3.2). This non-linearity often occurs in ecological data, and new methods of analysis were required to deal with it. Generalized Additive Models (Hastie and Tibshirani [1990]) were developed as part of these new methods. Rothery and Roy [2001] made use of GAMs when analysing BMS data (Chapter 2, Section 2.1.2).

GAMs are a nonparametric, smooth, data-led method of modelling, relying on the following assumptions:

- independence between observations,
- the model is additive (not multiplicative) in the covariates

3.10.1 Smoothers

Smoothers are a non-parametric method for allowing a relationship between the response variable y_{it} and the independent variables X_{it}. A simple example of this is the running mean smooth curve - at each data point y_i, the k nearest neighbours x_i are averaged, producing an estimate of the response. There are many variants of this, including Bin smoothers, kernel smoothers, running medians, and regression splines, all discussed in Hastie and Tibshirani [1990]. For all of these methods, the main issue is to find a balance between increasing the smoothness and increasing the goodness-of-fit to the data. This is a simple extension to linear models, Equation 3.2, which becomes:

$$Y_{it} = \sum_{j=1}^{p} f_j(X_{it}) + \varepsilon_{it}, \tag{3.22}$$

again, where $E(\varepsilon_{it}) = 0$, ε_{it} are independent, and $Var(\varepsilon_{it}) = \sigma^2$, and the f is a function of the covariates X to be estimated for p smoothing functions.

3.10.2 Regression Splines

A regression spline is a piecewise polynomial with breakpoints or join-points at knots, usually specified by the analyst. The more knots, the more flexible the curve is allowed to be - hence the term "spline", relating to the physical splines used by draftsmen to smooth a piece of wood between two points. Any increase in the knots (or related degrees of freedom) increases the flexibility (i.e., the "wiggliness") of the relationship between the covariates and the response. These splines, by definition, are smooth, with continuous derivatives allowing them to join smoothly at their knots.

B-splines are a form of regression spline, where the response y is regressed on an appropriate set of basis vectors. These basis vectors are functions representing some piecewise cubic polynomials. A covariate with one knot will be a matrix with four B-spline columns (one at each boundary region, an intercept term,

and one at the knot). Therefore, any covariate with K knots will have $K + 3$ columns in our framework. Knot placement can be selected using an automated data method, or by using prior biological or ecological information.

Generalized Additive Models (GAMs) are an extension of GLMs, still allowing a link function and a relationship between the covariates and the response (though, as above, this relationship does not have to be linear on any scale). The form of the model we will use here is similar to Equation 3.5:

$$E(Y_{it}) = \mu_i \tag{3.23}$$

and

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} f_j(X_{it}) \tag{3.24}$$

In the semi-parametric setting we will use, the $f(X_{it})$ can be either a linear or smooth term, dependent on the data.

3.11 Variance Inflation Factors

Before moving onto model selection criteria to choose between covariates, they should be tested for multi-collinearity (Fox [2002]). This occurs when two (or more) covariates are highly correlated. For example, butterflies might appear when there is a high sunshine percentage level and also when there are high temperatures. These covariates usually occur together and both would be used to the same effect. When multi-collinearity occurs, the matrix of covariates has strong dependence and may not be invertible, causing estimation problems. Even if estimation is numerically possible, the parameter estimates will be highly unstable. Variance Inflation Factors (VIFs) should be calculated before any model selection, in order to identify any collinearity. The VIF for the covariate j for $j = 1, ..., p$ estimated by β_j is calculated as follows:

$$VIF(\beta_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 here represents the correlation of the covariate x_j with all the other covariates (known as the multiple correlation coefficient). If the VIF is too big, usually the corresponding covariate is removed from any subsequent analyses. "Too big" is usually taken to be any value above 5.

Difficulty of multi-collinearity occurs when modelling with regression splines (Section 3.10.2), as the columns of the B-spline matrices are often highly correlated, which can cause parameter estimation problems. When comparing between smooth and linear terms, this potential problem should also be considered.

3.12 Model Selection

There are two main issues as regards model selection when using any of the above models - how to choose the "correct" correlation structure and how to choose the "correct" subset of covariates. Often these two issues become blurred, e.g., changing the correlation structure could lead to different standard errors and p-values for some covariates.

A third issue comes into play when dealing with GAMs, as described in Section 3.10.2. This concerns whether the relationship between the response (observed BMS count) and a covariate is linear or smooth, and if smooth, how many degrees of freedom to allow. Hastie and Tibshirani [1990] discusses some methods for investigating this issue.

As mentioned before, GEEs are not based on likelihood methods, and so any traditional model selection tools cannot be used - e.g., AIC or deviance as there is no model likelihood.

There is also a question of how to compare two competing models e.g., the GEE or the GAMM using the same covariance structure and subset of covariates. The models could be compared graphically (using partially fitted functions), or using a measure of goodness-of-fit such as deviance. Partial fitted curves here indicate the effect of a particular covariate on the response, whilst all other

	GLM	GLMM	GEE	GAM
AIC	✓	✓		✓
BIC	✓	✓		✓
QIC			✓	
% Deviance Explained	✓	✓		✓
Cross-validation	✓	✓	✓	✓
Likelihood ratio tests	✓	✓		✓
F-tests	✓	✓		✓

Table 3.1: Model Selection and Comparison criterion available for use for the different types of models.

covariates remain constant.

Table 3.1 summarises the model selection criteria which can be used for the different models.

3.12.1 AIC (and extensions)

AIC (Akaike [1973]) is the most commonly used model selection criterion for GLMs. It is based on the Kullback-Leibler discrepancy in decision theory and is defined to be:

$$AIC = -2l + 2p \tag{3.25}$$

where l is the log-likelihood for the model, and p is the number of parameters. It aims to choose the "best" subset of covariates by finding a trade-off between the improvement in fit due to including extra covariates and a penalty for adding extra terms - akin to Occam's Razor. Smaller values of AIC are preferred. AIC, however, cannot be used in the case of GEEs, as there is no likelihood involved (Section 3.8).

McQuarrie and Tsai [1998] and Anderson et al. [1994] suggest adding adjustment terms to the AIC statistic in order to adjust for bias caused by small sample sizes and time series data.

For example, with normal data:

$$AIC_c = AIC + \frac{2(p+1)(p+2)}{n-p-2} \tag{3.26}$$

with p regression parameters, and

$$CAIC = -2l + p[\ln(n) + 1] \tag{3.27}$$

Hurvich and Tsai [1989] show that using the AIC with auto-regressive time series data can lead to overfitting a model.

If overdispersion (Section 3.5) occurs, the AIC calculated may be unreliable, leading to unnecessary parameter inclusion.

There is no clear method which is best in all cases for dealing with overdispersion, though the unmodified AIC generally performs poorly.

The calculation of AIC also can become problematic when dealing with mixed models.

In Chapter 5, Section 5.3.2, the AIC statistic is presented for comparing some GAMMs (Section 3.9.1). The method used is based on a restricted pseudo likelihood (REPL) (Section 3.7), and the dispersion parameter is estimated afterwards. This is not ideal, and so AIC is not hugely dependable as a model selection criterion when comparing competing GAMMs.

AIC can be extended in two main ways. The first uses the traditional formula as in Equation 3.25, but allows the likelihood in the first term to be based on an extension to ordinary maximum likelihood; the second uses the traditional likelihood, but adds a term of adjustment.

There has been little research on model selection for autoregressive, overdispersed Poisson data.

3.12.2 Cross-Validation

Cross-Validation is another, data-based method of selecting covariates. Fewster et al. [2000] uses this method when choosing the "correct" number of degrees of freedom (see Section 3.10).

Generalized Cross-Validation (GCV) is often used with GAMs to choose between

competing smooth or linear fits, defined in the deviance setting as:

$$GCV = \frac{\frac{1}{n}\sum_{i=1}^{n} D(y_i, \hat{\mu_i})}{\left\{1 - \frac{tr(R)}{n}\right\}^2}, \tag{3.28}$$

where R is the weighted additive-fit operator (Hastie and Tibshirani [1990], Page 159) and $D(y_i, \hat{\mu_i})$ is the deviance of the model.
Cross-Validation is a data driven method of choosing between models.

3.12.3 QIC

As discussed above, the AIC is the most commonly used and powerful model selection criterion for any likelihood based models, but cannot be used for GEEs. A selection method, based on a modified AIC statistic is desired, and so the QIC method is introduced in Pan [2001]. For GEEs, a test which reflects the non-likelihood based aspect of the model is preferred. The QIC statistic can be used. The QIC is equal to the AIC when an independent model is selected (i.e., when counts on the same transect are assumed to be independent). The QIC uses quasi-likelihood. As with AIC, the lowest value of QIC is preferred.
QIC can also be used to choose the most appropriate correlation structure for the model (usually independent, exchangeable or first-order autoregressive).
QIC is defined as:

$$QIC(R) = -2Q(\underline{\hat{\beta}}(R); I, D) + 2trace(\hat{\Omega}_I \hat{V}_R)$$

for working correlation matrix R, where Q is the quasi-likelihood, as in Equation 3.16, $\underline{\hat{\beta}}(R)$ is the vector of maximum likelihood estimators under the candidate model with correlation structure R, I is the identity matrix, signifying an independence model, D is our observed data,

$$\hat{\Omega_I} = \frac{-\partial^2 Q(\underline{\beta}; I, D)}{\partial\beta\partial\beta'}|_{\beta=\hat{\beta}}$$

and $\hat{V_R}$ is the robust or sandwich covariance estimator estimated from the model containing working correlation matrix R.
Different forms for R can also be compared using this method.
For choosing the best subset of covariates, QIC is defined as QIC_u:

$$QIC_u = -2Q(g^{-1}(X\underline{\beta}(R))) + 2p \tag{3.29}$$

where Q implies using the quasi-likelihood which is calculated for the independence model, g^{-1} is the inverse link function of the model, and X is the matrix of covariates. Again, smaller values are preferred.

3.12.4 BIC

BIC, or Bayes information criterion (Hardin and Hilbe [2001]), is an extension of AIC, another likelihood based method, which is measured as:

$$BIC = -2l + p\ln(n) \tag{3.30}$$

where l is the log-likelihood of the model, p is the number of model parameters and n is the number of observations in the data set. The BIC is an approximation of the Bayes Factor, a statistic used in Bayesian statistics.
The BIC attaches a larger penalty to extra parameters in the model than AIC. For this reason, it is sometimes preferred, as modelers may prefer more conservative controls towards parameter inclusion.

3.12.5 F-tests

The F-test is generally considered to be a more conservative covariate inclusion test. It compares two nested models; the reduced model M_0 and the full model M_1 as follows in Equation 3.31:

$$F_{(0,1)} \frac{\frac{D_0 - D_1}{p_1 - p_0}}{\frac{D_1}{n - p_1 - 1}} \sim F_{(p_1 - p_0, n - p_1 - 1)} \tag{3.31}$$

This tests whether an extra parameter β_j is necessary in a model - a "large" F-value indicates that the extra parameter should be included. Obviously, this test is only valid when there is a valid deviance (i.e., a likelihood based model).

3.13 Offsets

It should be noted that sites with larger site area will inevitably have higher observed butterfly counts. It would lead to a misleading model to ignore this, and so site area should be included in the model as an offset. An offset is often used with data of this kind, where, for example, a patient will have a higher probability of catching a disease if they are exposed to it for a longer period of time. In this case, time exposed (to the disease) would be included as an offset. Since we are using the canonical Poisson log-link, $log(area)$ should be automatically included as follows:

$$\ln(E(y_{it})) = f(\{X_{it}\underline{\beta}\}) + \ln(area_i) \tag{3.32}$$

or

$$E(y_{it}) = \exp\{f(X_{it}\ \underline{\beta})\} \times (area_i) \tag{3.33}$$

for site i at time t, and $X{=}X_{it}$ is the matrix of model covariates.

3.14 Collating Indices

The most sensible and meaningful method used to collate indices is to "average" the site indices to produce a regional index. The two main methods of averaging are:

- to calculate the arithmetic mean ($I_R = \frac{\sum_{i=1}^{K} I_i}{K}$), and
- the geometric mean ($I_R = \prod_{i=1}^{K} I_i^{\frac{1}{K}}$),

where I_R is the calculated regional index, K is the number of sites, and I_i are the individual site-level annual indices. Moss and Pollard [1993] discuss these methods. Care must be taken when using the geometric mean, as zero values will distort and bias the results, however, this method gives each site equal weighting. The offset term has already accounted for site area.

Each mean has different properties - averaging by the arithmetic mean gives higher weighting to sites with higher abundances. Using the geometric mean essentially gives each site equal weighting, independent of abundance levels (Moss and Pollard [1993]).

Results using both these methods are presented in Chapter 5, Section 5.4.

3.15 Variance Estimation

It is well established in statistics that having an estimate of an amount is meaningless without an estimate of the variance of the estimate. As follows, we need to estimate the variance around our index estimates. We have demonstrated above different methods for calculating an (relative) index of butterfly abundance, either at site or regional level. The following sections summarise two methods for estimating confidence around these estimates. The first is a commonly used bootstrapping method (Section 3.15.1) and the second is a method used mostly in situations involving Mixed Models (Section 3.15.2).

3.15.1 Bootstrapping

Davison and Hinkley [1997] describe the basic idea behind bootstrapping. The main idea involves simulation based analysis, without relying on parametric model assumptions and analytical methods for calculating model uncertainty, when the data are more variable than a parametric model would suggest. With the advance of computer power, this method is easily implemented. The key idea involves resampling K sites with replacement from the K sites within a region, and re-estimating a regional index. This only produces a variance around

the regional estimate. Quantiles of the index (or any estimable function of it) can be easily calculated. Details of the method I implemented are in Chapter 4, Section 4.6.2.

Resampling is undertaken under the assumption that observations are independent, and for this reason, we resample whole sites rather than individual observations.

95% Bootstrapped confidence intervals can be calculated by taking the 2.5% and 97.5% percentiles. This generated interval is known as the Bootstrap percentile interval.

3.15.2 Variance-Covariance Method

The above method described in Section 3.15.1 relies on having powerful computers available for implementation. Often, this is unavailable to users, due to time or money constraints and so a quicker variance estimation method is also discussed here.

Mackenzie et al. [2005] describe a method used often in Mixed Models for estimating the variance of an estimate using the variance-covariance matrix.

This method involves simulating data from the multivariate normal distribution, with mean equal to the model parameters and variance equal to the variance of the model parameters.

This method differs philosophically from the bootstrapping method of Section 3.15.1 in that instead of resampling sites and estimating new models, this method assumes that a regional model is "correct" and simulates new data which can be used to generate confidence intervals around the "correct" indices. The traditional non-parametric bootstrap method (Section 3.15.1) has been found to underestimate the variability in the data, hence the variance-covariance method was preferred. There has been little work regarding using this method with GEEs, but it is useful for comparison's sake. This method can be used for calculating variances around site-level and regional-level indices.

As discussed in Section 3.9, the covariance structure of the Mixed Model is as

follows (Equation 3.34):

$$V(Y_i) = Z_i G Z_i' + \sigma^2 \tag{3.34}$$

where Z_i is a design matrix for the random effects parameters and σ^2 is the within-individual variability ($\varepsilon_i \sim (0, \sigma^2)$).

The variance around fixed effects parameter estimates are as follows (Equation 3.35):

$$V(\hat{\beta}) = K^{-1} V(Y) K \tag{3.35}$$

where K is a matrix chosen with as many linearly independent rows as possible and so that $K'X = 0$.

Assuming that the variance-covariance matrix structure is correctly specified, the matrix of fixed-effects parameter estimates is distributed as Multivariate Normal as follows (Equation 3.36):

$$\underline{\hat{\beta}} \sim MVT(E(\underline{\hat{\beta}}), V(\underline{\hat{\beta}})) \tag{3.36}$$

Values from this distribution can be simulated in order to create a data-set from which precision estimates can be calculated.

Analogously, values can be simulated from our selected GEE model where the the matrix of fixed-effects parameters are distributed as Multivariate Normal as follows (Equation 3.37):

$$\underline{\hat{\beta}} \sim MVT(E(\underline{\hat{\beta}}), V(\underline{\hat{\beta}})) \tag{3.37}$$

and

$$V(\underline{\hat{\beta}}) = K^{-1} V(Y) K \tag{3.38}$$

as above, but now $V(Y)$ is as in Equation 3.18.

In short, this approach generates predictions for daily butterfly abundance for each transect using the model coefficients and the corresponding variance-

covariance matrix. This is performed 1000 times and the 2.5th and 97.5th quantiles from these 1000 sets are used to give 95% confidence interval limits for each transect.

This method generates confidence intervals at site level, which can be combined to calculate regional-level intervals.

3.16 Comparing Indices Over Time

As the main topic of interest to wildlife managers and conservationists is the population change of the species over time, it is important to address this important issue. Methods are reviewed in Chapter 2, Sections 2.3.1 to 2.3.2, and the theory behind these methods is presented in the following sections.

3.16.1 Linear Route Regression

This very accessible model is based on the simple linear model of Section 3.3. The response (site-level annual index) is transformed by logging and site-level trends over time (years) are calculated as in Equation 2.5. A site level intercept and slope parameter are calculated. Trends are taken as the slope parameter β. Details vary between surveys, but many of these are presented in Thomas and Martin [1996]. Significance of trends can be tested using t-tests or z-tests.

3.16.2 Site by Years Model - Using TRIM

Similar to the above method described in Section 3.16.1, this method calculates regional-level trends using the software TRIM (Trends and Indices for Monitoring Data Pannekoek and van Strien [2005]). Using simple linear regression, as above, it transforms the annual regional indices by logging and calculates a separate parameter for each site and year (Equation 2.6).

TRIM also allows the inclusion of extra covariates.

An overall regional model is developed, with a separate parameter estimated for each site. This method can estimate regional level, as well as site level, trends.

3.16.3 Chain-Ratio Method

A traditional method used by the BMS and the UKCBC to compare indices over time is the ratio method (Cochran [1963]).

$$r_{t+1,t} = \frac{I_{t+1}}{I_t}, \tag{3.39}$$

where $r_{t+1,t}$ is the ratio between successive years, I_t is the regional index in year t and I_{t+1} is the corresponding index in year *t+1*. This method, and some extensions to it, is discussed in Mountford [1982]. One limitation of the method is that it is assumed that the same K sites are visited in each year.

The approximate variance ($V(r_t)$) of a Ratio estimate (Raj [1964]) is calculated as

$$V(\hat{r_t}) \doteq \frac{1}{[E(I_{t+1})]^2} V(I_{t+1} - \hat{r_t} I_t)$$

A change between two years t and *t+1* is found to be significant if r_t is significantly different from one. A confidence interval is calculated as follows:

$$\hat{r_t} \pm 2 \times (\frac{\sqrt{V(\hat{r_t})}}{K}) \tag{3.40}$$

If the confidence interval includes one, then there is no significant change in butterfly abundance between years t and *t+1*.

Confidence can also be calculated using a bootstrapping method (Section 3.15.1).

3.16.4 Discussion

There are many more methods which could be used in order to compare estimated indices over time - those presented in this thesis are just a small subset which have been used by organisations in the past. Results from using these methods on real BMS data are presented in Chapter 5.

3.17 Detection and Distance Sampling

In Chapter 2, Section 2.1.1, the issue of detection is introduced. The line transect method used by the BMS, described in Chapter 2, Section 2.1.2, only records the number of butterflies of each species. It is assumed that as many butterflies are missed as are recorded multiple times, and therefore, the number of butterflies recorded is taken as a proxy for abundance. This method does not address the problem of detection - that in more ideal conditions, independent of true abundance, the observed count will increase. Distance Sampling (Buckland et al. [2001]) is the method widely used to survey animals and calculate abundance and density. A transect line is walked in much the same way as the Pollard-Yates walk described earlier, though instead of recording sightings, distance to an observed object (i.e. butterfly) is recorded. These distances can be used to find a detection function. The estimation of the detection function depends on certain assumptions - these are:

- that $p(0) = 1$ - that detection at zero distance is certain,
- that objects (butterflies) do not move in response to the observer, and
- that distances are measured accurately.

$p(x)$ is the probability of observing an object at distance x from the transect line. One of the key assumptions in Distance Sampling is that every object on the line (i.e. at distance $x = 0$) is observed with probability 1 (i.e. is certain). Pollard [1977] describes a method which attempts to find a relationship between butterfly index and actual abundance. The method involves a mark-recapture study of 3 species at one site at Monk's Wood. The findings are encouraging, but involve too many assumptions and simplifications to be of much utility. As such, the BMS data we have can only lead to relative indices, and do not provide any information on abundances. This issue is re-examined in Chapter 4, Section 4.7.

Chapter 4

Statistical Methods used in the analysis of the Butterfly Monitoring data sets

4.1 Notation and Definitions

In this section, I will be using the following terms in the following context:

- y_{it}= count = observed (usually weekly) count made by an observer for site i at time (BMS week) t
- I= index =predicted butterfly (relative) index (at site-specific (I_i) or regional level (I_R)).
- $\Sigma_1^{26} y_{it}$= sum = sum of counts for site i and time t (BMS weeks) for $t = 1, ..., 26$ of the monitoring season.

4.2 Introduction

In this section I will describe the statistical methods used to calculate and compare butterfly indices (at site and regional level) using real BMS data collected. Variation around these indices is also discussed. Using these indices and variances, inference can be made on changes and trends in butterfly indices over time - which is one of the main aims of the BMS. Site level indices can be calculated by using simple addition of weekly counts (traditional BMS method) (Section 4.3.2), by modelling at site level (Section 4.3.3), or by developing a regional model to predict indices at site level (Section 4.3.4).
Field methods used in the BMS are described in Chapter 2, Section 2.1.2.

4.3 Calculating Annual (Relative) Indices of Abundance at Site-Level

4.3.1 Introduction

The aim of the BMS is to "produce an annual index of abundance at each site for all species recorded" - Rothery and Roy [2001]. In the following sections, I will describe different methods for producing these site-level indices using observed transect data.
Section 4.3.2 describes the original BMS method, which involves no modelling at all. It simply sums the weekly observations to produce a site-index, and uses linear interpolation to calculate missing values. If more than two consecutive weeks are missed, the site is not included. This method is clearly not utilising the data and information we have to its potential, and a more statistical approach is considered in the following sections.
Rothery and Roy [2001] make use of GAMs in order to fill in missing values. This GAM approach applies a flexible statistical model at site level. If no missing values exist on a particular site, this method simplifies to the original BMS method where weekly observations are summed to produce a site-level

annual index of abundance for a particular species. The strength of the GAM approach becomes evident when there are missing values in the data set (i.e., when a weekly count was missed). Instead of filling in gaps by simply averaging neighbouring observed counts, a smooth is applied, which aims to capture the flexible, smooth nature of the flight curve. This method still treats sites individually, and can be extended, as described in subsequent sections.

Often, sites within a geographical region can be very similar, and the flight patterns of butterflies on these sites can portray a very similar pattern. Instead of ignoring these similarities, we can utilise them to develop a regional model, which borrows strength and information from across the region in order to impute missing values and so predict indices at individual sites. There is a variety of methods which can be applied - these include GAMs, GLMs, GEEs, GLMMs and spatial models. Which of these is the most appropriate depends on the data and the question of interest. The different models can be compared using a variety of tests, as described in Section 4.3.5. Results are presented in Chapter 5.

4.3.2 Linear Interpolation Approach

The original approached used by the BMS to analyse data was to simply sum the weekly counts to produce a site level index of abundance. If a week was missed, linear interpolation was used to estimate missing values. If more than one consecutive week was missed, an index could not be calculated. This method is demonstrated on a small subset of our data.

4.3.3 Modelling Approach: Site level

Use of statistical models has greatly increased in recently years, due partly to the increased availability of computer power. The models utilise a robust, repeatable method, with theoretical properties, and so can be tested and compared. The most basic example of a statistical model is simple linear regression, which assumes a linear relationship between the site-level response (in our case

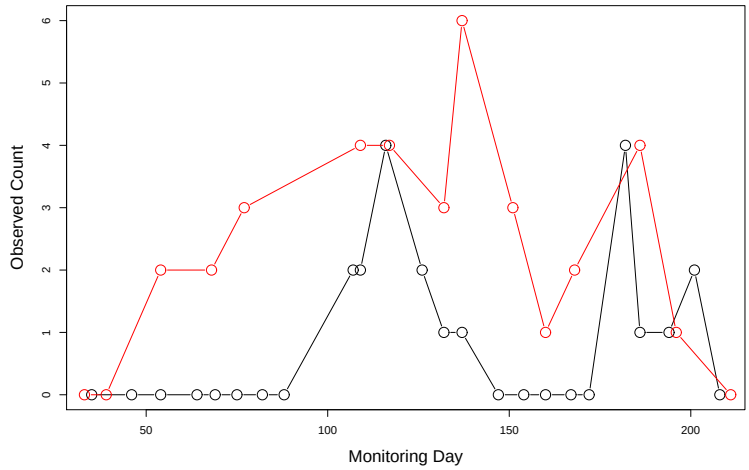


Figure 4.1: Observed Counts for sites 32 (in black) and 119 (in red) in the South Eastern Region, 2002, (Monitoring Day runs from March 1st to October 31st). Open circles indicate where observations were made.

- observed weekly butterfly counts) and the explanatory variable (monitoring week, or day). As can be seen clearly from Figure 4.1, it is not possible that a simple linear relationship could capture the flexibility and variability of the butterfly flight curve. Rothery and Roy [2001] use the flexible, smooth method of GAMs to model, at site level, the relationship between time and count.

GAM approach

As in Rothery and Roy [2001], this approach involves specifying a smooth model for each transect, used to predict missing values, and summing the 26 weekly counts as in Section 4.3.2. This can be interpreted as an estimate of the area under the flight period distribution curve. Technical details of GAMs are found in Chapter 3, Section 3.10.

4.3.4 Modelling Approach: Regional level

Neighbouring sites produce very similar flight curves (e.g., see Figure 4.1). In order to reliably fill in missing site-level data, especially when more than one consecutive week is missed, we can make use of regional level information. A

regional level model can be developed, using information from all sites surveyed, and the model is subsequently used to make predictions at individual site level. There are many different ways of capturing the flexibility of the flight curve and describing the similarities and differences between neighbouring sites. Some of these models include GAMs, GEEs, GLMMs and spatial models. These are described in more detail in the following sections, along with methods of comparing these models.

4.3.5 Model Comparison

"Model comparison" is used in place of "model identification", as we have no way of knowing or identifying what the true number of butterflies at any site is or what the true underlying model is. There are many different statistical tests which can be used to compare subsets of covariates within any particular model, as in Chapter 3, Section 3.12. However, it is a more difficult question to compare different overall models. Our task is to use sensible judgement in applying these tests, given the underlying assumptions of each model, and given the variability captured using each model. I aim to produce summary statistics for each model considered, as appropriate, and also to use graphical methods to compare models.

Before applying any of the following models to the data, model covariate selection was undertaken in order to prevent any unnecessary parameters being estimated. A step.gam function was run on each data-set to choose between inclusion of terms as linear or smooth functions, or dropping them from the model. Knot selection was investigated, with results in the following chapter.

4.3.6 GAMM approach

The aim of this section is to describe the mixed model approach for these data. A GAMM was fitted to the South East, 2002 data and the results compared with the GAM/GEE approach. The extra complexity of this approach (compared with the GAM/GEE method) was not found to be justified and the non-random

way in which these data were obtained meant a mixed model approach was philosophically problematic. For these reasons, the GAM/GEE-based model was preferred to the GAMM approach.

Generalized Additive Mixed Models assume the underlying model holds for each transect inside each region, but allow model coefficients to vary across transects and/or regions. Underlying this modelling framework is the assumption that the sites sampled represent a random sample from a larger population of sites. Mixed model analysis requires at least one term be considered as random in addition to the more traditional (fixed) terms. In this analysis, 'BMSDay' was specified to be random to allow the indices of abundance to vary with the day of the season across transect. A random intercept effect was also considered in order to allow baseline abundances to vary across transects. The model contained the following covariates after covariate selection: 'Habitat type', 'BMSDay', 'Temperature', 'Sun', 'Wind', 'Easting', 'Northing', 'Altitude' and 'Time of Day'.

4.3.7 Spatial Model Approach - allowing for more flexible Temporal Correlation across Regions

The aim of this section was to investigate the merits of fitting a model to a subsection of the data which allows for the temporal correlation to vary depending on location. Specifically, the extent of correlation in the counts within transects over time was allowed to vary across counties instead of assuming a constant regional-level correlation coefficient. The penalised fit measure for this model suggested this additional complexity was not justified and a regional-level measure for the temporal autocorrelation was preferable.

A spatial model was fit to the data using the data from the South Eastern region, 2002. This was performed to see if spatial information needed to be considered, over and above the temporal autocorrelation accommodated by the GAM/GEE approach. Unfortunately, convergence for a model which explicitly models spatial autocorrelation in the data could not be obtained. As an

alternative, the coefficient for the temporal AR(1) correlation was permitted to differ spatially in the attempt to accommodate additional spatial variation. The covariates fitted in the model were the same for each model type, as above: Habitat type, BMSDay (day of year on which count was made),Temperature, Wind-Speed, Altitude, Easting and Northing.

4.4 GEE Approach

4.4.1 Aim

The aim of this section is to describe a GEE approach to analysing the BMS data and calculating site-level indices. Technical details are provided in Chapter 3, Section 3.8. Regression splines (Section 3.10.2), as described, are extremely useful in describing the non-linear butterfly flight patterns. However, GAMs assume that all observations are independent and as demonstrated in Chapter 3, Section 3.8, correlation within transects should be considered. An approach which allows the flexibility of GAMs along with allowing for correlation is considered and discussed in the following section.

4.4.2 Model Specification

For any given site, successive counts are likely to be related to one another and this correlation is unlikely to be described by the model in full. Further, fitting models (i.e. GAMs or GLMs) which ignore this correlation can result in unreliable measures of precision about abundance estimates. As an alternative, a GEE-based model which allows the response (i.e. count) to be modelled as a function of covariates can accommodate this correlation. A Poisson GEE for the South East was fit containing nine candidate covariates. Counts per unit area were modelled (using an area 'offset') and the errors, ε_{it}, were assumed to

have an AR(1) structure and were permitted to be overdispersed:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat_i + f_1(BMSDay_{it}) + f_2(temp_{it}) + f_3(sun_{it}) + f_4(wind_{it}) + f_6(east_i) + f_7(north_i) + f_5(alt_i) + f_8(time_{it})\} \tag{4.1}$$

where y_{it} correspond to the counts for site i at time (monitoring day) t, $f_1, ..., f_8$ represent smooth terms for each covariate. Regression splines were used to model the smooth functions and added flexibility was permitted for 'BMSDay' to allow for multiple brooding. A covariance structure must be chosen for the GEE model and GEE results can also be used to test whether counts within transects are independent. For this analysis, the AR(1) structure was used to model the non-independence. This structure allows dependence amongst successive counts, with the dependence decreasing as counts are further apart in time.

At this stage, it is worth mentioning that instead of taking BMS Week as the unit of time at which an observation was made, we take BMSDay. This is to allow for greater precision and smoother flight curves. We impute counts for every day, which produces a smooth flight curve as in Figure 4.2. Site level indices were previously calculated by summing the weekly counts. We integrate under the flight curve, which is essentially equivalent to summing the predicted daily counts for every day in the Butterfly Monitoring season.

4.4.3 Model Selection

Model selection was completed for each of the different regions in turn (South East, Anglia, East Midlands, North East, Scotland, South Central, South West, Thames and Wales). It was first completed for the South Eastern region, 2002 in the following way:

Variance inflation factors (Chapter 3, Section 3.11) were calculated for covariates within this region. Once no intolerable levels of multi-collinearity was found

for these data, all candidate variables were considered for model selection.

A Poisson GAM/GEE was fitted containing the nine potential covariates. Counts per unit area were modelled (using an area 'offset') and the errors, ε_{it}, were assumed to have an AR(1) structure and were permitted to be overdispersed, just as in Equation 4.1.

Tests were performed in order to test whether each variable was required in this full model (in a smooth form, i.e. as $f(BMSDay_{it})$) and a small p-value indicates a variable was included.

However, a problem exists when considering covariate selection when using regression splines. For example, for the BMSDay covariate, there are 5 terms to be included. p-values calculated using available software test these 5 terms individually (using, e.g., Wald tests), and they should be tested as a group for inclusion (either as splines or as linear terms).

QIC was also used, although it is not yet an automated method.

4.4.4 Code

As this method (of using regression splines within a GEE, allowing for model selection) is not automated, code in R was written as follows (see Appendix A for code).

Steps in the program are as follows:

- Data checking for variance inflation factors (VIFs)
- Model selection
- Evaluating daily predictions
- Bootstrapping for standard errors and confidence intervals.

Details are presented here, and results are presented for the Small heath species, for the South East region, 2002 data set in Chapter 5, Section 5.3.4. The aim of running this program is to take all of the observed counts, over different transects within a region, and develop a regional model to describe the flight

curve of the butterfly species of interest (in this case, Small heath). Then, using this model, predictions are made for every transect for every day in the butterfly monitoring season , using median site-level covariates. This produces a smooth curve for every transect, and daily counts are summed to produce a site level index of abundance, using the regional model.

Firstly, variance inflation factors are checked to be sure that collinearity will not be present as a problem. Covariates with VIFs greater than five are removed at this stage. Variable selection takes the following steps:

- A GAM is run, with all potential covariates included.
- From this, a step.gam function is run to choose between linear and smooth terms, or omitting terms, using AIC for the stepwise selection.
- A step function is then run using BIC, in order to be stricter on covariate inclusion.
- The covariates selected are then used in the regional model to estimate daily counts at transect level.

For this particular data set, the steps are run as follows: If none of the VIFs were found to be greater than five, the step.gam included all potential covariates. The step.gam function chooses between smooth and linear terms for BMSDay, sunshine, temperature, wind, east, north, altitude and time of day. Habitat-type is also tested for inclusion as a factor.

The subsequent step function also chooses between models - stricter inclusion rules might reject some covariates.

This model was used to produce site level indices as described in the next section. For every transect, the model is used to predict a count for days in the monitoring season (from day 20-230, in this case). Median covariates are chosen for every site - i.e., for transect 32, sunshine=100%, temperature $=20 \circ C$, wind-speed=2, northing=1450, easting=6070, altitude=100 and time of day=290 (minutes after 9am). A plot for transect 32 of day against counts, also with predictions

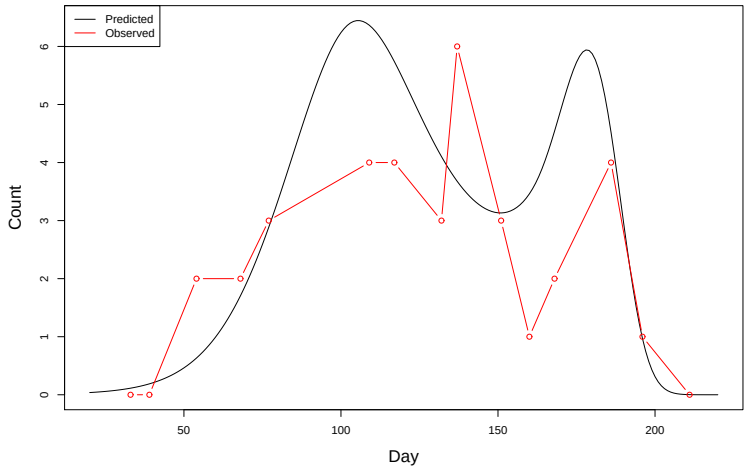


Figure 4.2: Observed and Predicted Daily Counts for Transect 119, 2002. Monitoring Day runs from March 1st to October 31st. Predicted counts are in black, observed counts in red.

made is in Figure 4.2. These graphs show how the linear graph, with missing values, is replaced by a smooth curve.

Predictions are made for each day in the Butterfly Monitoring Season (BMS), unlike the traditional linear interpolation method (Section 4.3.2), which just sums the weekly observed counts. Since we have predicted a smooth site-level flight curve, it makes more sense, and is more precise, to integrate fully under the curve, which is equivalent to summing the daily predictions.

The issue of detection (Chapter 3, Section 3.17) also needs addressing. Covariates such as percentage sunshine will invariably have an impact on species detectibility. Predicting at standard covariates over time would perhaps be more correct when comparing indices over time. Whether predicting at previously standardised covariates or at median site-level covariates makes any practical difference shall also be investigated.

Results after applying this method to Small heath data are contained in Chapter 3, Section 5.3.5.

4.5 Collating Site-Level Indices to produce Regional Indices

Knowing that a species is increasing or decreasing at any specific site does not give us information about the wider countryside. A method is needed to collate the site-level indices to produce regional indices. As previously mentioned, though, the non-random nature of the sites surveyed mean that even regional indices produced do not guarantee information about the greater countryside area.

In this section, I will describe different methods used to produce these regional indices.

4.5.1 Simple Addition

The simplest method is to sum the site level indices to produce regional indices. This method will produce an annual regional index, however, it is not scaled for the number of sites included in any region.

4.5.2 Arithmetic Mean

An extension to the summation method is the arithmetic mean. Site level indices are summed, and the result divided by the number of sites in the region. $I_R = \frac{\sum_{i=1}^{K} I_i}{K}$, where I_i is the annual index for site i, for the K sites surveyed within the region.

4.5.3 Geometric Mean

Moss and Pollard [1993] suggests that the geometric mean is the recommended method of collating individual indices: $I_R = \prod_{i=1}^{K} I_i^{\frac{1}{K}}$

4.6 Variance of Indices

As always when using statistical methods, it is important to produce variance estimates for any analyses. In this section, I will describe different methods of calculating variances around indices - both site level and regional level.

4.6.1 Site level - using Model Variance-Covariance matrix

The selected regional GEE/GAM model is used to predict indices at site-level. Using this model, and its associated covariance matrix of the model parameters, precision of these indices can also be estimated. The function "rmvnorm" in the Statistical Package "R" is used 1000 times to draw random values from the multivariate normal distribution with mean equal to the model parameters and covariance matrix sigma, where sigma is the covariance matrix of the model parameters. These random variables are used as model parameters to make daily predictions as before.

As always, daily predictions are summed to give an annual site-level index of abundance. As 1000 such index values are generated, confidence intervals and levels of precision can be calculated.

95% confidence intervals can be generated by ranking the 1000 indices and picking out the 2.5% and 97.5% values.

4.6.2 Regional Level

Again, it is more meaningful to have precision of regional level indices rather than those for site level. Traditional bootstrapping can be applied, as well as combining the site-level precision described above.

Bootstrapping

Fewster et al. [2000] suggest using a bootstrap method. K sites are selected with replacement from the K sites surveyed for a given year, and for these K sites, a regional model is selected, which predicts indices for every site. This process is

repeated B times (usually, $B = 1000$) and so B regional indices are calculated using a method from those listed in section 4.5. The 2.5th and 97.5th percentiles of these B indices are found, and so a 95% confidence interval is calculated for the annual regional index.

Combining Site-level variance

Site-level precision can be calculated as in section 4.6.1. The method can be extended to produce regional precision as follows. The 1000 index values for each site can be collated to produce 1000 regional indices (e.g., the B=1st indices for each site are averaged to produce the first of 1000 regional index). Confidence intervals are calculated in the usual way.

4.7 Comparing Indices over Years

Again, the main aim of the BMS is to make inferences about butterfly populations over time. By calculating indices and having precision estimates of these indices, we can discuss any changes over time. The following section describes the methods used in this thesis.
These methods are applied to some data in the following chapter, and results are presented in Section 5.6.

4.7.1 Confidence Interval of Differences

There are $B = 1000$ bootstrapped indices calculated annually for each region. The difference between each $B = 1, ..., B = 1000$ sample between successive years is calculated, and a 95% Confidence Interval of differences found. If the confidence interval (of differences) includes zero, no difference is assumed. Conversely, if the confidence interval does not contain zero; this indicates a change in (relative) butterfly index between years at the 5% level of significance.

4.7.2 Ratio Method

The ratio method is described in Chapter 3, Section 3.16.3.

Between successive years, only sites which are present in both years are included in the analysis. These sites are used to select regional-level models as described in Section 4.4.4, and site-level predictions are made using standardised covariates (a temperature of $20C°$, wind-speed of 2, % sunshine of 80% and time of day of 13.10). Regional-level collated indices are calculated, and subsequently ratios between successive years, as in Equation 2.8.

These sites can then be bootstrapped as before, and B bootstrapped indices calculated. 95% confidence intervals can be generated, and intervals not including one indicates a significant difference between years.

Chapter 5

Results of Analyses on the BMS data

The first part of this chapter describes the data used in the following sections. As stated previously, the main aim of the Butterfly Monitoring Scheme is to compare Butterfly indices over time, at site, regional and national levels. Sections 5.2 to 5.5.2 of this chapter show results from different methods used to calculate site-level and regional-level indices. Section 5.6 aims to compare these estimated indices over time.

5.1 Data Description

The Small Heath butterfly, *Coenonympha pamphilus* (as discussed in the following section), is the multivoltine species being considered here. It has very variable phenology, having up to 3 broods per season, especially in southern regions, and so requires a very flexible method of modelling to capture this.
There are data from 1976 to 2002, covering twelve regions of the United Kingdom (Anglia, East Midlands, North East, Northern Ireland, North West, South Central, Scotland, South East, South West, Thames, Wales and the West Midlands) - in all, 316 data sets.

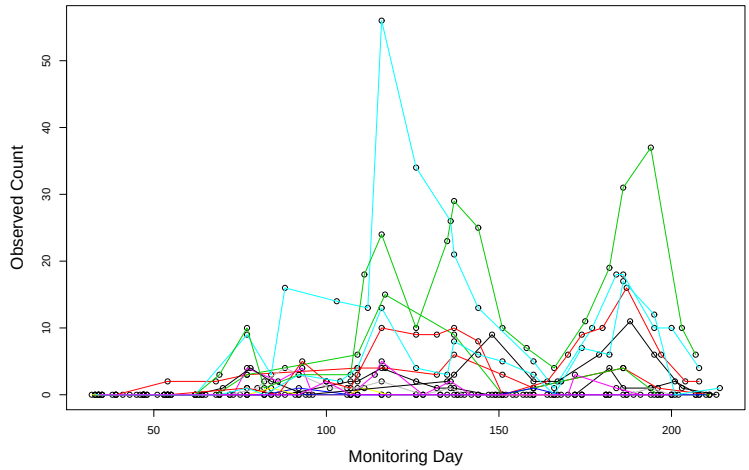


Figure 5.1: Weekly Observed Counts for all transects for South Eastern Region, 2002. Monitoring Day runs from March 1st to October 31st.

A simple plot of the South East region for 2002 (Figure 5.1) demonstrates the variability in counts for a double brooded area. Figure 5.2 shows a corresponding single brooded area - Scotland in 1988. Each colour represents a different site within a region - in all, there were 14 different transects surveyed in the South East in 2002. The South East region covers the counties of Kent, Surrey, Sussex, Greater London, London and West Sussex. Habitats 1,3 and 7 are covered - these represent Native Woodland, Calcareous Grassland scrub and Various (parkland, mixed, upland, etc.) in the BMS broad habitat descriptions. Any number from 3 to 30 transects were surveyed annually in the South East over the 27 years, with a mean of 13.5 and a median of 14. As described in Chapter 1, transects were walked weekly from March to September (the butterfly monitoring season). Ideally, each transect should be walked 26 times, but as explained previously, there are many missing values. The weekly, observed counts range from 0 to 56, with a mean of 3.1 and a median of 0. This tells us that the data are very right skewed, as is seen in Figure 5.3. Butterfly counts across sites were highly variable (Figure 5.4) ; most sites have very few counts, while transect 71 has some high counts and transect 2011 has a particularly

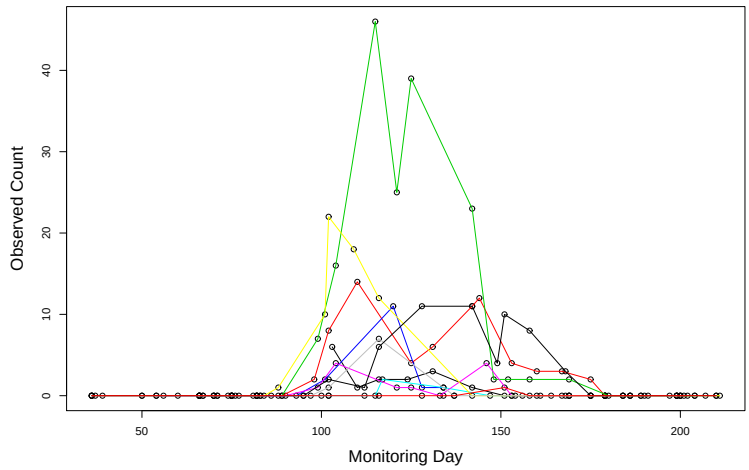


Figure 5.2: Weekly Observed Counts for all transects for Scotland, 1988. Monitoring Day runs from March 1st to October 31st.

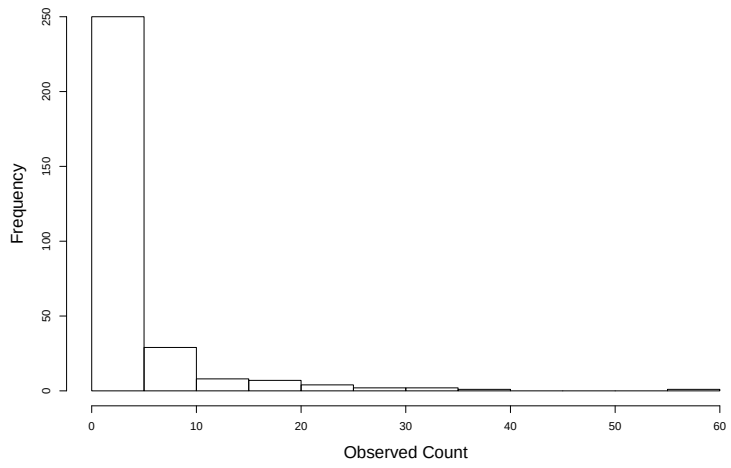


Figure 5.3: Histogram of Observed Counts for the South East Region, 2002

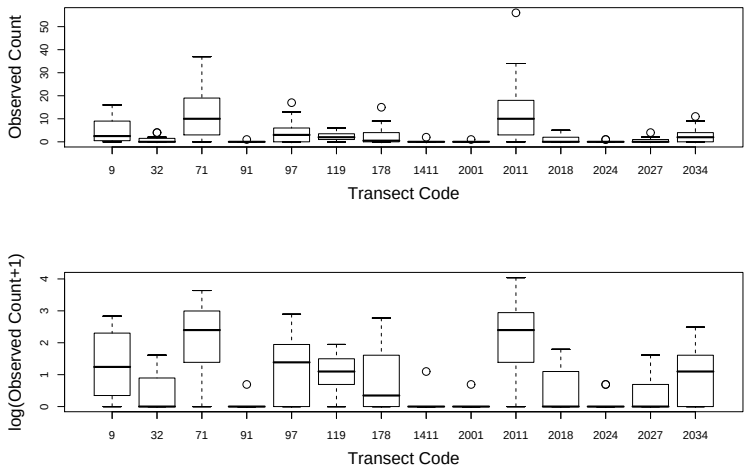


Figure 5.4: Boxplots of counts by Site, and logged counts by site for the South East, 2002 data.

high count of 56. As the data are most complete for the South East, these are the data analysed in the following sections, but the methods work well on any other data sets.

For various reasons, some surveys finished before the end of the monitoring season. This caused instability in the data, and hence in the models considered. An example is in the data for the South Central region, 2002, see Figure 5.5. To deal with this problem, zero counts were added to the beginning and end of each data-set in order to anchor the data and represent a time when no butterflies are expected. Figure 5.6 shows data from the South Central region, 2002, with the extra anchored data added in. To be consistent, these zeroes (two at each end) were added to every data-set, at days 1, 8, 239 and 245, corresponding to March 1st, March 8th, October 25th and October 31st. These data were collected by the CEH and BC.

5.1.1 Small Heath Butterfly

The Small Heath butterfly (or *Coenonympha pamphilus*) is a small but conspicuous, widespread resident UK butterfly. It is a relatively abundant, sedentary

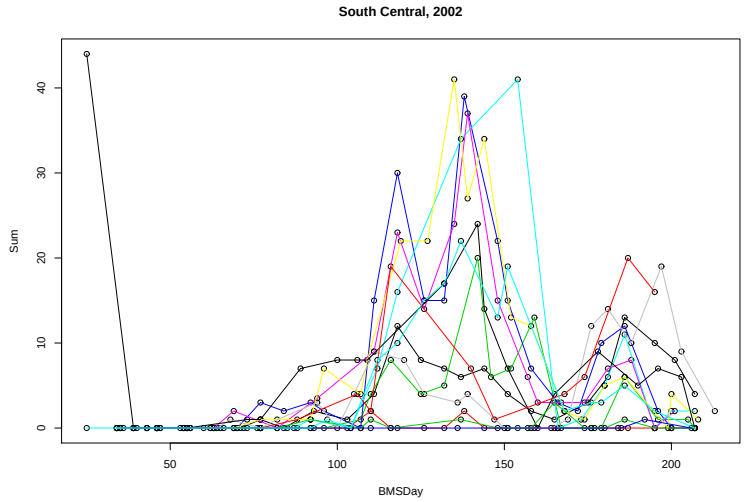


Figure 5.5: Weekly Observed Counts for all transects for South Central, 2002. Monitoring Day runs from March 1st to October 31st.

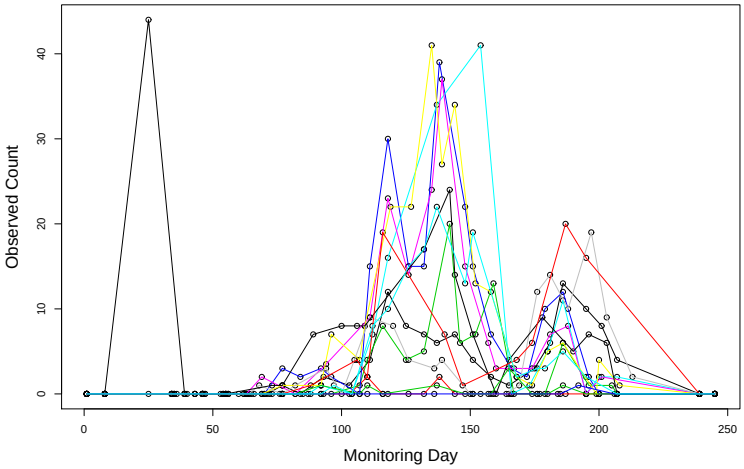


Figure 5.6: Weekly Observed Counts for all transects for South Central, 2002, with anchored zeroes at days 1, 8, 239 and 245. Monitoring Day runs from March 1st to October 31st.

butterfly, strongly associated with open grassland habitat, although found occasionally in most habitat types. It is a greyish brown colour, and mostly found resting on grasses. It is a member of the Satyridae family, which includes the Large Heath, Meadow Brown, Ringlet and Speckled Wood. It is very commonly found in Britain and Europe, and has been sighted as far south as North Africa and as far east as Mongolia. It is not on any threatened species lists, though its biggest threat is habitat-loss. It has a complex life history, with sometimes up to 3 broods per year. The first brood usually appears in April or early May, with the second in August. With increasing temperatures, especially in southern England, it commonly has 3 broods per year, with usually only one in Scotland. Since the data for this species are quite abundant, it has been the focus for the analyses that follow.

5.1.2 Explanatory Variables

Both site-specific and time-varying covariates were potentially included in the model.

These are:

- BMSDay - the day on which each survey took place, from BMSDay 1 = March 1st to BMSDay 245 = October 31st,
- temp - the temperature , in C°, at the site when the survey took place,
- wind - the wind-speed (from 1 to 5 on the Beaufort Scale), at the site when the survey took place - any speeds higher than 5 were considered too windy for the survey to proceed,
- % sunshine (from 0% to 100%) at the site when the survey occurred, though scaled to be a number between 0 and 10,
- easting and northing - Ordinance Survey Grid reference of the site,
- alt - altitude of the site (in metres),

- time - start time of the survey (minutes past 9am),
- habitat - habitat type of the site - codes 1 to 7 derived for the Butterfly Monitoring Scheme:
  - 1 = Coastal
  - 2 = Bog, Moor and Wetland
  - 3 = Grassland, Bracken and Scrub
  - 4 = Woodland
  - 5 = Farmland
  - 6 = Urban, Industrial
  - 7 = Other
- area - site area is used as an offset in the model (in metres squared).

Site area was calculated as transect length multiplied by transect width (usually 5 metres). Details of how the BMS habitat classifications correspond to other schemes (i.e., EUNIS cross-referencing) are provided in Appendix B.

5.2 Calculating Annual (Relative) Indices at Site Level - Linear Interpolation Approach

This is the traditional method used by the BMS to calculate abundance indices. Table 5.1 shows results for the South East region, 2002, using this method, described in Chapter 4, Section 4.3.2 It can be seen that for 5 out of the 14 transects surveyed in 2002, an index cannot be calculated due to too many missing values. This method is far too simplistic and involves no modelling. The following sections will describe more robust methods which deal with missing values using more sophisticated statistical techniques.

Transect Name	Transect Code	Missing weeks	Index
Wye & Crundale Down	32	2 (2,10)	21
Folkestone Escarpment	119	11 (3,5,8,9,10,11,14,17,21,22,25)	NA
Cheriton Hill	178	12 (3,5,8,9,10,11,14,15,17,21,22,25)	NA
Banstead Downs	2001	4 (2,7,10,18)	1.5
Denbies Landbarn	2011	4 (2,6,10,18)	309.5
Juniper Hill, Walton Downs	2018	2 (25,26)	24
Oaken Wood	2024	2 (23,25)	2
Park Downs	2027	1 (2)	14
Whitedown	2034	7 (1,2,10,11,12,14,18)	NA
Kingley Vale	9	2 (5,6)	NA
Castle Hill	71	0	284
Woods Mill	91	6 (2,10,11,14,15,18)	NA
Lullington Heath	97	1 (10)	97.5
Bevendean A	1411	1 (26)	2

Table 5.1: Indices for the South East, 2002, by site, calculated using the traditional BMS linear interpolation method.

5.3 Calculating Annual (Relative) Indices at Site-Level using a Regional Level Model

This section develops a single regional model using information from all sites within the region. This model is then used to predict counts for each individual site, using site-specific information. As has been seen in Figure 4.1, sites within a region can be very similar, so it makes sense to borrow strength across sites to develop this regional model. There are many approaches that can be taken here, as described in the following sections.

5.3.1 Covariate Selection

Before comparing different models (GEEs, GAMs, GAMMs), covariates need to be selected. Covariate selection specifically for GEEs is discussed in Section 5.3.4.

The step.gam function in R was run on the selected data and the following

model was chosen:

$$E(y_{it}) = \exp\{\beta_0 + f(BMSDay_t) + f_1(sun_{it}) + f_2(temp_{it}) \\ +f_3(wind_{it}) + f_4(east_i) + f_5(north_i) + f_6(alt_i) + time_{it}\} \times (area_i) \tag{5.1}$$

i.e, habitat is not found to be necessary for this particular data-set, start-time is included as a linear term, and all other covariates are included as smooth terms.

These selected terms in Equation 5.1 are then used for comparing competing model types (GAMMs and GEEs).

Knot selection

Knots for the smooth terms need to be specified. Knots should be placed where there is sufficient data to support them. The median of each covariate is often used, but in this case, the median was often at one of the extremes of the data-set (i.e., the median for % sunshine was 100%, which is also the maximum). This caused problems of convergence, and so the mean of each covariate was used instead. This could potentially be an issue, as sometimes the mean is not actually a data-point, but it did not cause any problems in any of the Small heath BMS data analysed. An extra knot was allowed for the smooth BMSDay term, allowing for extra flexibility due to the two peaks - the knots were chosen at days 116 and 175 (June 24th and August 22nd).

5.3.2 GAMM Approach

The term GAMM refers to a GLMM (Chapter 3, Section 3.9.1) with smooth, additive terms (Chapter 3, Section 3.10) allowed. The GAMM approach I used assumes that the underlying model is over-dispersed Poisson, with covariates included as in Equation 5.1. Two GAMM models were considered:

- the first allowing a random intercept term (i.e., assuming that β_0 is not fixed, but that $\beta_0 \sim N(0, G)$), allowing the baseline abundances to vary

Model	AIC	ϕ (p-value)	ρ (p-value)
Over-Dispersed GLM	1672.0	4.096 (< 0.0001)	NA
Random Intercept	1645.4	4.156 (< 0.0001)	0.1637 (0.0083)
Random Day	1645.4	4.156 (< 0.0001)	0.1637 (0.0083)

Table 5.2: Summary Statistics for the two GAMMs considered, and those for the corresponding Overdispersed GLM model.

across sites,

- the second allowing the parameters for BMS Day to vary across the region - allowing the day of peak emergence to vary across sites, as sometimes occurs with real data.

Fitted curves and model summary statistics were compared for these two models considered.

Philosophically speaking, we assumed here that:

- the sites surveyed were a random, representative sample taken from the region, and
- the days were randomly surveyed.

Obviously, neither of these assumptions are valid when considering the real BMS data sets - sites were self-selected either by wildlife managers or local volunteers as sites with a known, established butterfly presence, and surveys occurred prominently on sunny weekend days - not randomly chosen.

Convergence was not reached for the model with random terms for all 5 BMS-Day splines, and so only the 3 middle terms were allowed to be random - this is where most of the flexibility is needed.

Table 5.2 presents the AIC statistic (adjusted for dispersion), dispersion-parameter ϕ and correlation coefficient ρ (along with p-values) for the two generalized additive mixed models considered, along with those corresponding to the over-dispersed GLM approach. Figure 5.7 shows partial fitted curves for BMSDay for the GAMMs, and the over-dispersed GLM. Partial fitted curves here indicate the effect of a particular covariate (in this case, BMSDay) on the response, whilst

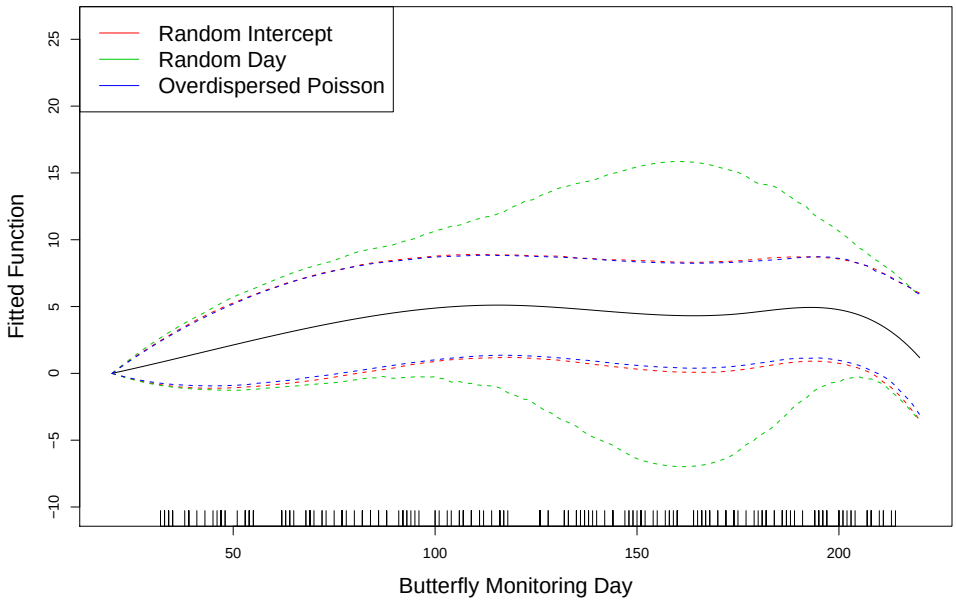


Figure 5.7: Partially Fitted Functions for BMSDay for the South Eastern Region, 2002, with 95% Confidence Intervals for the models with Random Intercept, Random BMSDay and over-dispersed Poisson GLM. Partial fitted curves here indicate the effect of a particular covariate (in this case, BMSDay) on the response, whilst all other covariates remain constant.

County	Correlation Estimate	p-value
Kent	0.107	0.441
Surrey	0.144	0.147
Sussex	0.000	1.0
Common structure	0.1034	0.075

Table 5.3: AR(1) Correlation Parameter Estimates, with p-values, for the different Counties in the South East Region, 2002, and also allowing a Common Correlation structure for the Region.

all other covariates remain constant. Maximum likelihood (ML) is used for obtaining summary AIC statistics and Restricted Maximum likelihood (REML) is used for the variance estimation, as discussed in Chapter 3, Section 3.9.2.

Code for the mixed models was run in SAS - the GLIMMIX macro command fits mixed models and "Proc Transreg" allows for smooth terms.

SAS Code is presented in Appendix C.

5.3.3 Spatial Model Approach

Previous models allow for correlation within transects, but assume that all the transects within a region display the same levels of correlation (i.e., one common ρ is estimated across the region). In this section, I allowed the correlation in an AR(1) GEE model to vary across counties, with a separate correlation parameter ρ calculated for each (in the South East, 2002, there were three surveyed: Sussex, Kent and Surrey).

The correlation in the counts across time at the transect level appeared to be strongest in Surrey, and less pronounced for Kent and no correlation evident for Sussex (Table 5.3, Figure 5.8). More specifically, there was reasonable evidence for within transect correlation in Surrey, weak evidence of this correlation in Kent and no evidence of within transect correlation in Sussex. Regardless of the correlation estimates across counties, the correlation within counts over time was estimated to drop to approximately zero after 5 weeks (i.e., counts further than 5 weeks apart were no longer dependent under the model). The fit statistics under each model suggested that correlation coefficients at the county

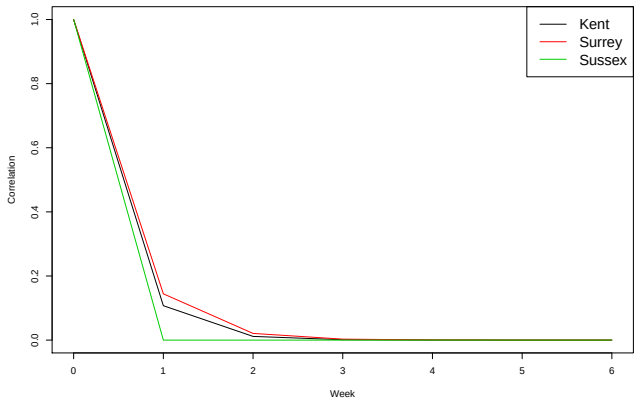


Figure 5.8: Correlation between counts (AR-1) for South East Counties, 2002

Model	AIC
Common Correlation	1645.4
County Level Correlation	3182.6

Table 5.4: AIC Summary Statistics for Comparing a Spatial Model allowing for County Level Correlation and a Model allowing only a Common Regional Level Correlation Structure, for the South East data, 2002.

level were not justified; the AIC when correlation was considered at the county level was 3182.6 compared to 1645.4 for a common correlation structure across counties (see Table 5.4). Also, there was very little difference in the precision in the fitted curves when the correlation structure was assumed to be common to all counties or varying across counties (Figures 5.9 to 5.11).

5.3.4 GEE Approach

GEE Model Selection

The GEE approach begins with all covariates included as smooth terms, and the independent correlation matrix. QIC is used then to compare models:

- between smooth and linear terms for numerical covariates,
- between omitting any of these terms, and
- between independent and other (as described in Chapter 3, Section 3.8.2)

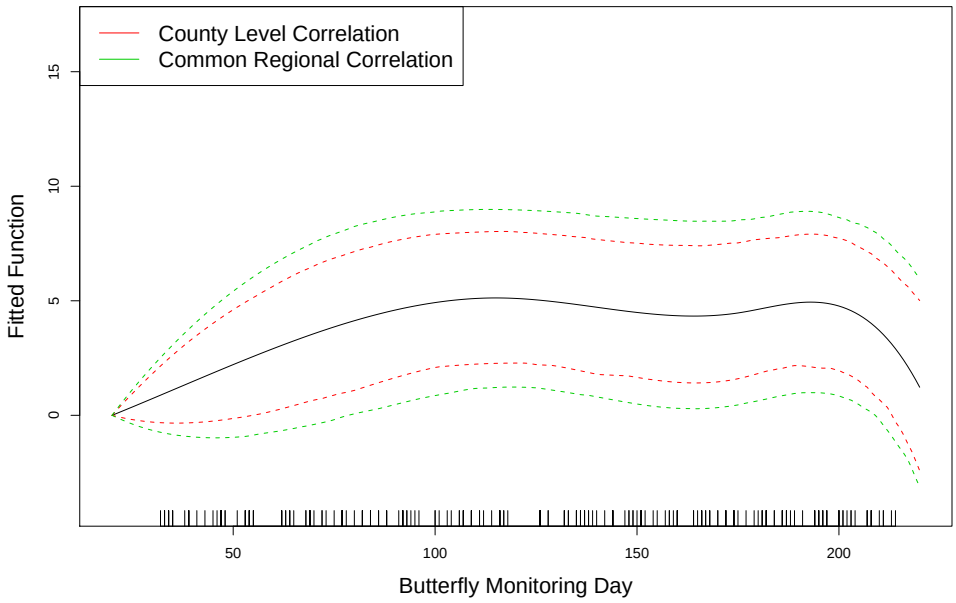


Figure 5.9: Fitted curves for BMSDay for a model with a common correlation structure across counties (in green) and a model with correlation structure which is allowed to vary across counties (in red). These data were collected in the South Eastern Region, 2002.

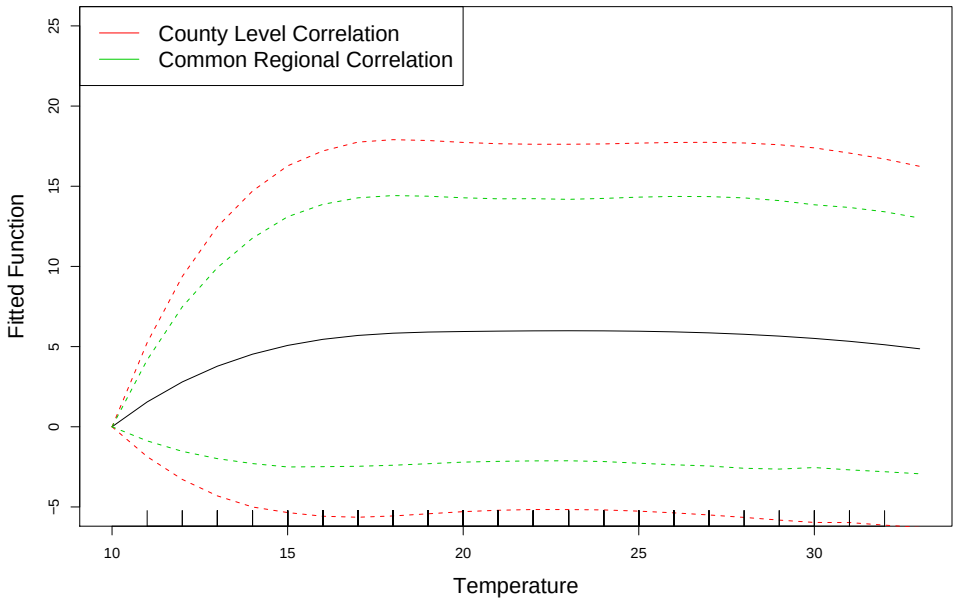


Figure 5.10: Fitted curves for Temperature for a model with a common correlation structure across counties (in green) and a model with correlation structure which is allowed to vary across counties (in red). These data were collected in the South Eastern Region, 2002.

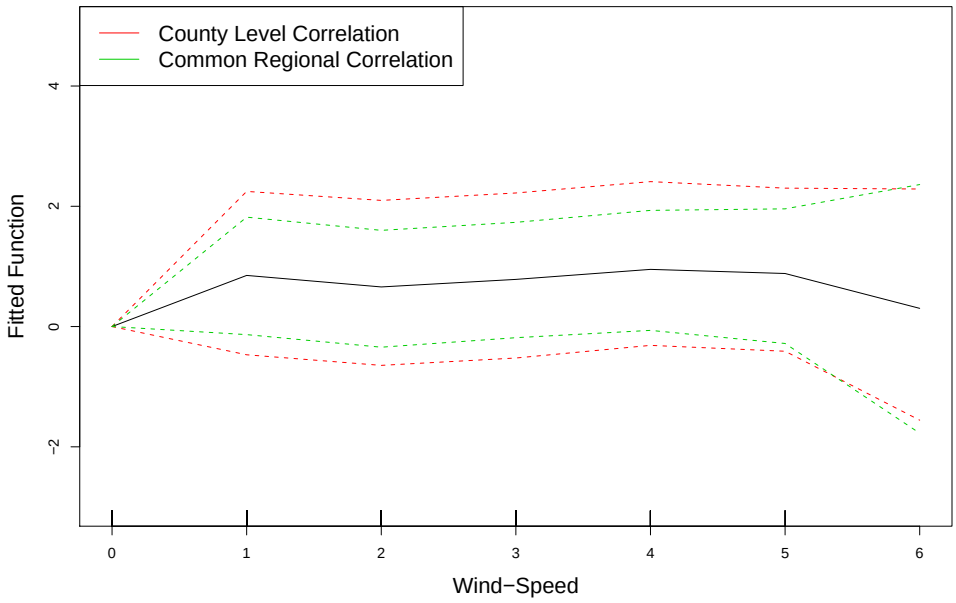


Figure 5.11: Fitted curves for Wind for a model with a common correlation structure across counties (in green) and a model with correlation structure which is allowed to vary across counties (in red). These data were collected in the South Eastern Region, 2002.

Model	QIC - Ind	QIC - AR(1)	QIC - Exch
Full	492.8	492.7	491.9
Linear Day	736.3	733.6	737.3
Without Day	755.6	751.3	754.8
Linear Temperature	508.0	509.3	506.6
Without Temperature	505.6	506.9	504.4
Linear Sun	499.7	501.7	497.1
Without Sun	514.1	515.8	513.2
Linear Wind	498.7	499.6	498.8
Without Wind	506.8	509.0	507.4
Linear East	791.4	862.9	796.7
Without East	790.6	848.8	794.4
Linear North	1667.7	NaN	1738.7
Without North	1582.9	NaN	1587.8
Linear Altitude	590.1	588.8	589.7
Without Altitude	589.0	587.9	588.5
Linear Time	489.5	486.8	488.5
Without Time	493.9	490.7	493.0
Linear Time + Habitat	NA	NA	NA

Table 5.5: QIC values for different GEE models under consideration for the South East data set, 2002.

correlation matrices.

Results lead us to the following model:

$$E(y_{it}) = \exp\{\beta_0 + f_1(BMSDay_t) + f_2(sun_{it}) + f_3(temp_{it}) \\ + f_4(wind_{it}) + f_5(east_i) + f_6(north_i) + f_7(alt_i) + time_{it}\} \times (area_i) \tag{5.2}$$

Details of model comparison using QIC is given in Table 5.5, where "Ind" refers to an independent model, "AR(1)" refers to first order auto-regressive correlation, and "Exch" refers to exchangeable correlation. The "Full" model is taken as Equation 5.2. Any further model comparison leads to the same model being selected using the QIC criterion, as in Table 5.6, where the "Full Model" is the model chosen from Table 5.5. Habitat type can be included as a factor, and tested using an F-test. The F-test statistic of habitat inclusion (based on an overdispersed GLM) is calculated to be 0.0020. This indicates that habitat type is not necessary in the model (Table 5.7).

Model	QIC - Ind	QIC - AR(1)	QIC - Exch
Full	489.5	486.8	488.5
Linear Day	734.9	727.9	736.1
Without Day	758.4	749.5	758.6
Linear Temperature	505.2	504.7	503.0
Without Temperature	502.8	502.6	501.1
Linear Sun	497.3	495.7	494.8
Without Sun	511.7	510.7	511.0
Linear Wind	494.8	492.6	494.9
Without Wind	502.5	501.2	503.2
Linear East	790.6	836.0	797.2
Without East	794.4	827.6	798.5
Linear North	1670.4	NaN	1742.0
Without North	1585.1	NaN	1742.0
Linear Altitude	589.5	588.5	589.3
Without Altitude	589.7	588.8	589.4

Table 5.6: QIC values for further GEE models under consideration for the South East data set, 2002.

Model	F-statistic	p-value
Full model (with habitat)		
Reduced model (without habitat)	0.002	1

Table 5.7: F-test of inclusion of habitat as a factor in the over-dispersed GLM model with regression splines for the South East data, 2002.

Transect Name	Transect Code	Index
Wye & Crundale Down	32	171
Folkestone Escarpment	119	594
Cheriton Hill	178	511
Banstead Downs	2001	20
Denbies Landbarn	2011	2062
Juniper Hill, Walton Downs	2018	183
Oaken Wood	2024	16
Park Downs	2027	95
Whitedown	2034	758
Kingley Vale	9	816
Castle Hill	71	1701
Woods Mill	91	3
Lullington Heath	97	831
Bevendean A	1411	9

Table 5.8: Indices for the South East, 2002, by site, calculated using the GEE method with Regression Splines (to the nearest Butterfly), predicted at site-median time-varying covariates.

GEE indices

Once a model has been chosen, as above, predictions are made for every site and every day in the Butterfly Monitoring Scheme, using the site-median for each time-altering covariate, from day 20 to day 220 (March 20th to October 6th, these days are the minimum and maximum for our data). The flight curve is smooth, and a selection are presented in Figures 5.12 (for the South East) and 5.13 (for Scotland). The flight curve is integrated (i.e., summed) to give a site-level annual index of abundance for the Small heath butterfly. Results are presented for the South East, 2002, in Table 5.8, though results presented without precision estimates are not very meaningful, and so are presented in the next sections.

As discussed, to be comparable over time, indices should be predicted at a standard set of time-varying covariates. Whether this makes a practical difference to predictions is considered at a later stage. Results comparing the two methods for regional-level indices for the South East region are presented in Table 5.9.

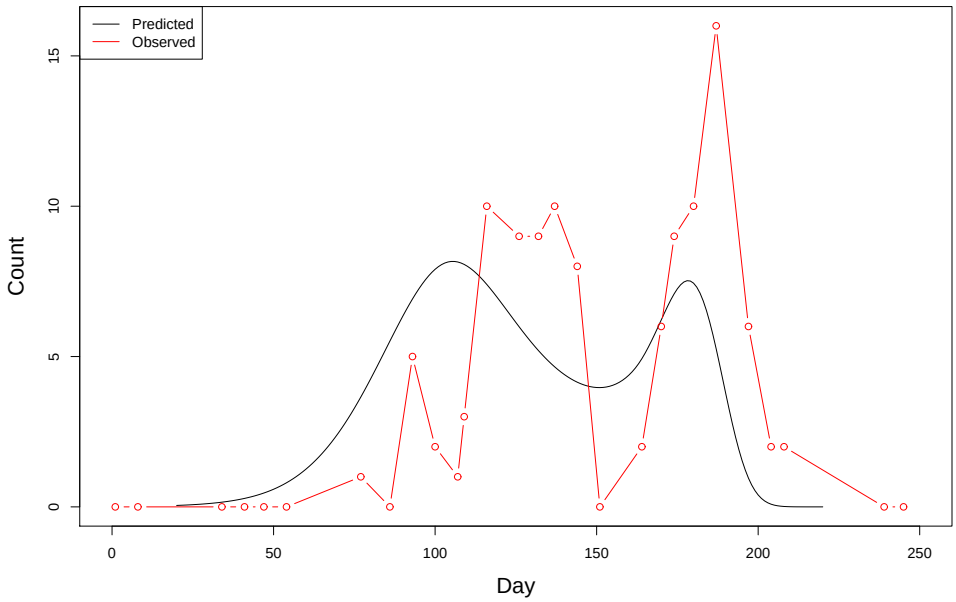


Figure 5.12: Observed and Predicted Daily Counts for Transect 9, South East, 2002. Monitoring Day runs from March 1st to October 31st. Predicted curve is in black, observed counts in red.

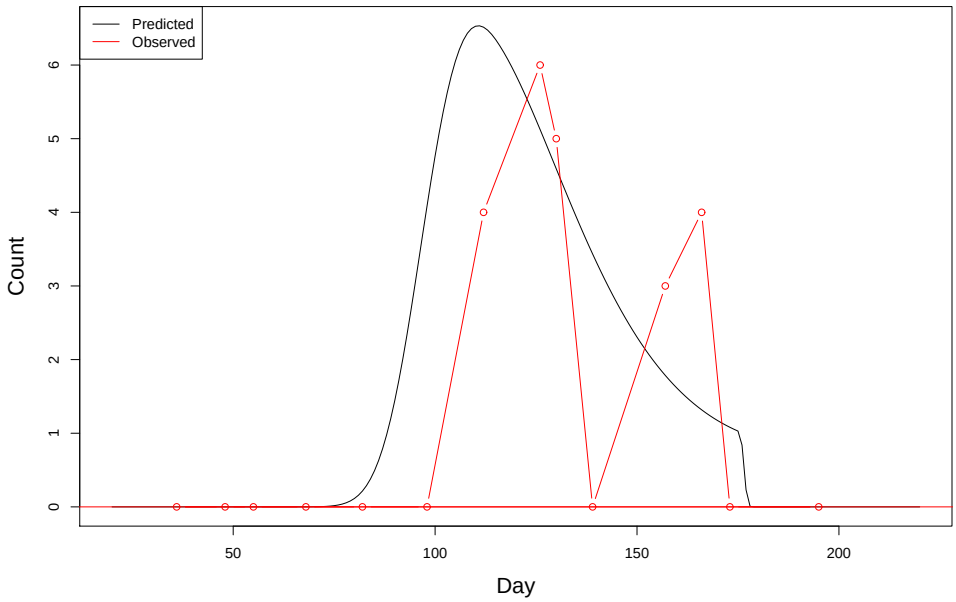


Figure 5.13: Observed and Predicted Daily Counts for Transect 61, Scotland, 2002. Monitoring Day runs from March 1st to October 31st. Predicted curve is in black, observed counts in red.

Year	A. Mean (Med.)	G. Mean(Med.)	A. Mean (Std.)	G. Mean (Std.)
1977	1636	176	473	143
1978	677	93	873	142
1979	1929	1067	3118	1872
1980	1705	340	1251	273
1981	684	217	712	241
1982	1338	584	1090	493
1983	1311	402	1000	324
1984	1621	378	1239	331
1985	583	184	511	190
1986	1115	313	993	295
1987	1039	292	912	324
1988	772	238	1146	350
1989	2281	631	1981	560
1990	1518	751	1858	849
1991	1057	376	1007	351
1992	775	36	1058	42
1993	420	145	420	162
1994	703	153	636	179
1995	1174	324	787	281
1996	2094	382	1696	361
1997	1390	232	1257	209
1998	553	373	683	437
1999	401	268	301	200
2000	216	140	114	75
2001	361	197	461	237
2002	555	169	548	159

Table 5.9: Regional Indices for the South East, 2002, calculated using the GEE method with Regression Splines, predicted at site-median time-varying covariates (Med.) and at standard covariates (Std.) (to the nearest Butterfly).

5.3.5 Other Regional Results

The above automated method was run on data for all regions and GEE models selected as follows:

For Anglia, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + sun_{it} + east_i + north_i + time_{it}\} \tag{5.3}$$

For East Midlands, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + sun_{it} + f(wind_{it}) + f(time_{it})\} \tag{5.4}$$

For North East, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + sun_{it} + f(wind_{it}) + (time_{it})\} \tag{5.5}$$

For Northern Ireland, 2002:

The model did not converge, as only one transect was surveyed, and so there was not sufficient data to support a GEE model.

For North West, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + temp_{it} + f(sun_{it}) + f(time_{it})\} \tag{5.6}$$

For South Central, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + f(sun_{it}) + f(wind_{it}) + f(east_i) + f(north_i) + f(time_{it})\} \tag{5.7}$$

For Scotland, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(sun_{it}) + east_i + north_i + f(time_{it})\} \tag{5.8}$$

For South West, 2002:

No model was found for the South West Region, 2002, as there was not sufficient data to support a GEE model.

For Thames, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + f(sun_{it}) + f(wind_{it})\} \tag{5.9}$$

For Wales, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + f(sun_{it}) + f(wind_{it}) + north_i + alt_i\} \tag{5.10}$$

For West Midlands, 2002:

$$E(y_{it}) = area_i \times \exp\{\beta_0 + habitat + f(BMSDay_{it}) + f(temp_{it}) + sun_{it} + f(wind_{it})\} \tag{5.11}$$

Correlation coefficients and dispersion parameters for all the other regions are presented in Table 5.10, along with corresponding p-values. Results are also available for all other years. Negative correlation parameters imply patterns of negative correlation. This is cause for concern, and occurs only in regions with less than 6 sites annually surveyed. Negative correlation is estimated along with very high p-values, implying that correlation is not evident at such sites.

Region	No. sites	ϕ (p-value)	ρ (p-value)
Anglia	8	3.65 (0.020)	0.016 (0.8525555)
East Midlands	5	2.88 (0.010)	0.800 (0.008)
North East	4	0.879 (0.0005)	-0.017 (0.893)
Northern Ireland	1	NA	NA
North West	4	190.05 (0.999)	0.0004 (0.999)
South Central	13	3.00 (0.107)	0.266 (0.162)
Scotland	6	0.789 (0.863)	-0.031 (0.858)
South West	3	NA	NA
Thames	4	3.10 (0.905)	0.275 (0.934)
Wales	7	2.23 (0.003)	0.051 (0.502)
West Midlands	4	NA	NA
South East	14	1.739 (0.839)	0.239 (0.835)

Table 5.10: Correlation coefficients (ρ) and dispersion parameters (ϕ) (with p-values), from Regional-level models selected, for all Regions surveyed in 2002.

5.4 Collating Site-Level Indices to produce Regional Indices

The theory behind the following methods is presented in Chapter 3, Section 3.14.

5.4.1 Simple Addition

This is the method traditionally used for BMS analyses, but are meaningless as over time, sites are not always surveyed, and so different numbers of sites are surveyed annually, and so these indices are not comparable.

5.4.2 Arithmetic and Geometric Means

Regional indices are calculated for Scotland and the South East region using the arithmetic and geometric means over all sites surveyed in each region. Results are presented in Tables 5.11 and 5.12.

Year	# Sites	A. mean	95% CI	G. mean	95% CI
1980	8	154	(16,1.7e+82)	79	(0,164)
1981	9	335	(40,883)	142	(19,285)
1982	11	1072	(138,3985)	207	(35,903)
1983	10	814	(59,2204)	187	(20,618)
1984	10	637	(299,1352)	421	(78,838)
1985	12	501	(237,1168)	322	(112,589)
1986	11	137	(34,415)	59	(18,141)
1987	9	157	(52,3027)	90	(8,443)
1988	10	507	(180,2159)	170	(60,608)
1989	9	1614	(312,13552)	557	(32,2472)
1990	8	777	(258,1612)	447	(68,939)
1991	7	1561	(179,42770)	440	(21,3507)
1992	11	462	(57,4.7e+38)	154	(29,714)
1993	9	64	(11,685034)	25	(1,308)
1994	11	193	(61,4536)	91	(8,386)
1995	9	930	(90,42095718)	196	(52,4198)
1996	14	429	(158,2089)	199	(42,588)
1997	12	1436	(215,63853)	237	(69,1084)
1998	12	591	(141,3.1e+10)	0	(0,5607)
1999	15	819	(203,2032)	383	(96,937)
2000	14	368	(113,5017)	204	(28,922)
2001	16	277	(132,914)	111	16(365)
2002	6	228	(59,Inf)	118	(36,Inf)

Table 5.11: Indices for the Scotland Region over time, calculated using the GEE method with Regression Splines, with 95% Confidence Intervals. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of 1000 bootstrapped samples, bootstrapped by site. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly).

5.5 Precision Estimates

5.5.1 Bootstrapping by site

Regional indices are presented for Scotland and the South East in Tables 5.11 and 5.12. These include bootstrapped 95% confidence intervals. These indices are calculated using site-median time-varying covariates. Graphs of these indices over time for the South East are presented in Figures 5.16 and 5.17, and for Scotland in Figures 5.14 and 5.15.

[1]For Figure 5.15, year 2002 and Figure 5.14, years 1991, 1992, 1993, 1995, 1997, 1998 and 2002 are unbounded due to upper bounds which are too large.

[2]For Figure 5.17, years 1990, 1995 and 2000 are unbounded due to zero values multiplied into the geometric mean.

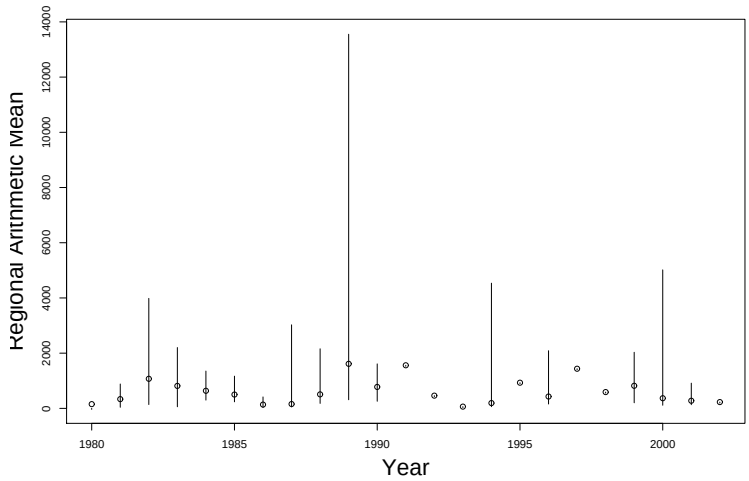


Figure 5.14: Arithmetic mean Regional Indices for the Scotland Region over time, with 95% Confidence Intervals, calculated by bootstrapping sites.
1

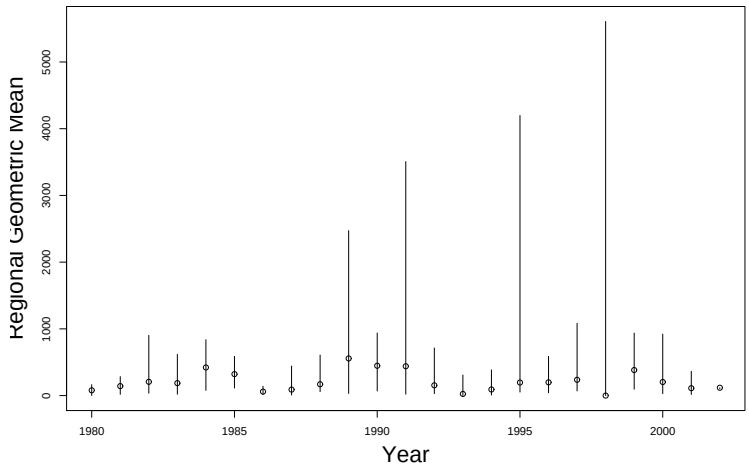


Figure 5.15: Geometric mean Regional Indices for the Scotland Region over time, with 95% Confidence Intervals, calculated by bootstrapping sites.
1

Year	# Sites	A. mean	95% CI	G. mean	95% CI
1980	6	1705	(344,2928)	340	(56,1018)
1981	6	684	(142,2817)	217	(17,891)
1982	6	1338	(444,3742)	584	(27,2023)
1983	8	1311	(459,4120)	402	(52,1720)
1984	9	1621	(378,3790)	378	(75,1692)
1985	9	583	(152,1303)	184	(34,526)
1986	8	1115	(400,3058)	313	(89,1984)
1987	9	1039	(823,3067)	292	(288,1558)
1988	10	772	(188,1996)	238	(27,1089)
1989	15	2281	(758,4025)	631	(96,1531)
1990	17	1519	(804,2987)	751	(NA,NA)
1991	17	1057	(432,2279)	376	(112,833)
1992	19	775	(427,1503)	36	(0,686)
1993	15	420	(176,996)	145	(46,407)
1994	14	703	(208,1463)	153	(37,449)
1995	14	1174	(386,2240)	324	(NA,NA)
1996	16	2094	(801,4209)	382	(92,1186)
1997	19	1390	(823,2036)	232	(101,853)
1998	25	553	(301,1331)	373	(198,665)
1999	25	401	(191,873)	268	(109,435)
2000	29	216	(96,649)	140	(NA,NA)
2001	23	361	(148,735)	197	(76,381)
2002	14	555	(259,1122)	169	(29,538)

Table 5.12: Indices for the South East Region over time, calculated using the GEE method with Regression Splines, with 95% Confidence Intervals. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of 1000 bootstrapped samples, bootstrapped by site. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly).

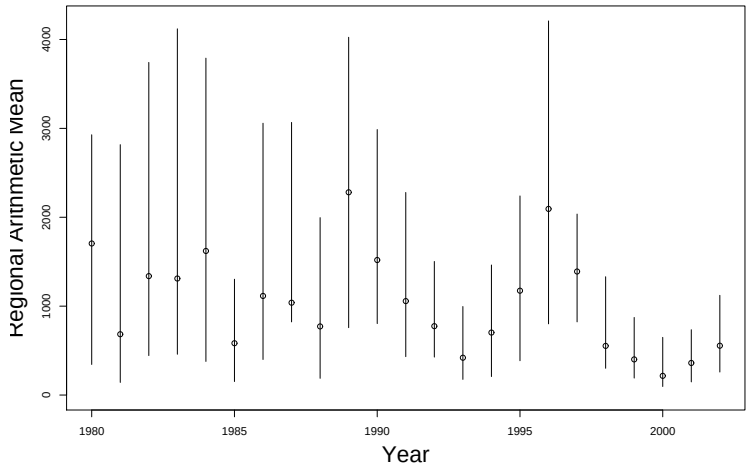


Figure 5.16: Arithmetic mean Regional Indices for the South Eastern Region over time, with 95% Confidence Intervals, calculated by bootstrapping sites.

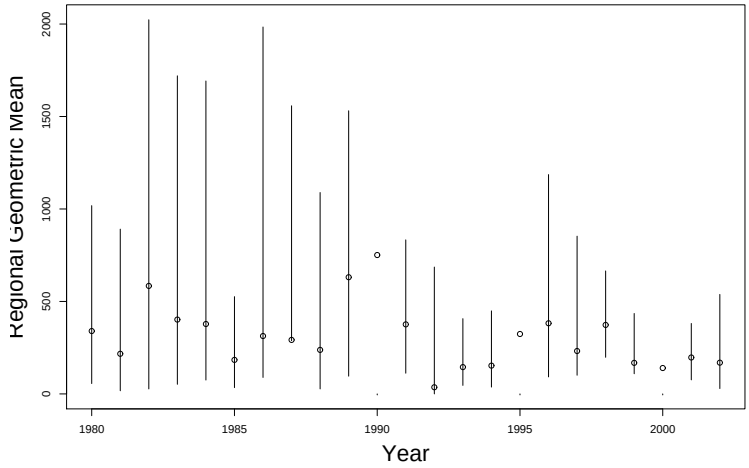


Figure 5.17: Geometric mean Regional Indices for the South Eastern Region over time, with 95% Confidence Intervals, calculated by bootstrapping sites.
2

Year	# Sites	A. mean	95% CI	G. mean	95% CI
1993	15	420	(266,11401637)	145	(89,242)
1994	14	703	(382,601403)	153	(59,10246)
1995	14	1174	(924,5423)	324	(382,924)
1996	16	2094	(1854,4306)	382	(481,2094)
1997	19	1390	(1402,4.5e+29)	232	(1,638041)
1998	25	553	(421,953)	373	(252,537)
1999	25	401	(322,1049)	268	(223,537)
2000	29	216	(150,2752)	140	(67,358)
2001	23	361	(213,1148)	197	(87,552)
2002	14	555	(526,2.6e+43)	169	(0,147868)

Table 5.13: Indices for the South East Region over time, calculated using the GEE method with Regression Splines, with 95% Confidence Intervals. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of 1000 simulated values using the Variance-Covariance matrix. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly).

Code used for bootstrapping sites is presented in Appendix E.

5.5.2 Using the Variance-Covariance Matrix

Regional indices are presented for a subset of years for the South East in Table 5.13. These include 95% confidence intervals calculated by simulating from the Variance-Covariance Matrix.

Obviously, the predictions are the same using the two methods, however, there are discrepancies in the values of the confidence intervals. Code for this method is presented in Appendix D.

5.6 Comparing Indices over Time

5.6.1 Differences between bootstrapped indices

Indices are compared over time using the simple confidence interval method. The difference between 1000 bootstrapped indices for a region over two years was calculated. Next, the 2.5th and 97.5th percentiles of the difference was noted. This formed a 95% confidence interval of the difference between successive years. If the confidence interval included zero, no difference was noted. However, if the

Years	Diff. A. mean	95% CI	*	Diff. G. mean	95% CI	*
(1980-1981)	-181	(-790,Inf)		-63	(-233,87)	
(1981-1982)	-737	(-3802,441)		-65	(-772,164)	
(1982-1983)	258	(-1614,3733)		20	(-521,716)	
(1983-1984)	177	(-1109,1762)		-235	(-740,317)	
(1984-1985)	136	(-659,1022)		99	(-310,575)	
(1985-1986)	364	(-20,1042)		263	(26,525)	
(1986-1987)	-20	(-2891,236)		-30	(-377,85)	
(1987-1988)	-350	(-1665,2690)		-80	(-525,301)	
(1988-1989)	-1107	(-13872,918)		-387	(-2126,253)	
(1989-1990)	837	(-733,13493)		109	(-563,2053)	
(1990-1991)	-784	(-42669,678)		8	(-3280,599)	
(1991-1992)	1099	(-442,42745)		286	(-424,3342)	
(1992-1993)	398	(-3.8e+7,1170)		129	(-186,692)	
(1993-1994)	-129	(-4270,37875547)		-67	(-305,224)	
(1994-1995)	-737	(-4.2e+07,2809)		-105	(-4006,204)	
(1995-1996)	501	(-1366,64725424)		-3	(-421,3951)	
(1996-1997)	-1007	(-63095,1112)		-38	(-822,335)	
(1997-1998)	845	(-1.5e+10,51265)		237	(-5183,832)	
(1998-1999)	-228	(-1366,1.5e+10)		-383	(-843,5085)	
(1999-2000)	451	(-4554,1662)		179	(-531,716)	
(2000-2001)	91	(-473,4749)		93	(-250,774)	
(2001-2002)	49	(-Inf,598)		-7	(-Inf,235)	

Table 5.14: Differences (with 95% Confidence Intervals) between annual Regional BMS indices for the Scotland Region. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of the differences between 1000 bootstrapped samples, bootstrapped by site. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly

confidence interval of differences did not include zero, then a change in butterfly abundance across years was accepted. The results for the years 1977 to 2002, for the regions Scotland and the South East are presented in Tables 5.14 to 5.16, where a * indicates a difference between years.

5.6.2 Ratios between Bootstrapped Site-level Indices

Using the method described in Chapter 4, Section 4.7.2, 1000 bootstrapped indices were calculated for the years between 1980 and 2002 in the South East Region. Results are presented in Table 5.17. In order to be comparable across years, predictions were made at a standard set of time-varying covariates across the region. Again, a * indicates a significant difference between years (i.e. the

Years	Diff. A. mean	95% CI	*	Diff. G. mean	95% CI	*
(1980-1981)	1021	(-2247,2237)		123	(-597,869)	
(1981-1982)	-654	(-2989,2161)		-367	(-1770,661)	
(1982-1983)	27	(-3283,2640)		182	(-1478,1701)	
(1983-1984)	-310	(-2777,2908)		24	(-1285,1466)	
(1984-1985)	1038	(-384,3302)		194	(-278,1457)	
(1985-1986)	-532	(-2695,394)		-129	(-1827,250)	
(1986-1987)	76	(-2021,1798)		21	(-1219,1335)	
(1987-1988)	266	(-724,2312)		54	(-509,1378)	
(1988-1989)	-1508	(-3373,709)		-393	(-1349,715)	
(1989-1990)	762	(-1533,2513)		-120	(-1138,894)	
(1990-1991)	462	(-977,2079)		375	(-308,1212)	
(1991-1992)	281	(-701,1519)		340	(-416,735)	
(1992-1993)	355	(-245,1155)		-109	(-358,510)	
(1993-1994)	-283	(-1087,489)		-8	(-311,289)	
(1994-1995)	-471	(-1732,679)		-171	(-918,329)	
(1995-1996)	-920	(-3330,873)		-58	(-1015,771)	
(1996-1997)	705	(-739,2947)		150	(-584,872)	
(1997-1998)	836	(-176,1520)		-141	(-442,534)	
(1998-1999)	152	(-339,979)		106	(-141,466)	
(1999-2000)	184	(-323,630)		128	(-83,325)	
(2000-2001)	-144	(-500,315)		-58	(-287,117)	
(2001-2002)	-194	(-838,341)		28	(-368,269)	

Table 5.15: Differences (with 95% Confidence Intervals) between annual Regional BMS indices for the South Eastern Region. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of the differences between 1000 bootstrapped samples, bootstrapped by site. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly

Years	Diff. A. mean	95% CI	*	Diff. G. mean	95% CI	*
(1993-1994)	-283	(-482296,11399836)		-8	(-10136,96)	
(1994-1995)	-471	(-3691,598536)		-171	(-646,9676)	
(1995-1996)	-920	(-2936,3020)		-58	(-1494,201)	
(1996-1997)	705	(-4.5e+29,1109)		150	(-636316,1626)	
(1997-1998)	836	(749,4.5e+29)	*	-141	(-457,637700)	
(1998-1999)	152	(-533,460)		106	(-211,253)	
(1999-2000)	184	(-2311,771)		128	(-46,407)	
(2000-2001)	-144	(-759,2253)		-58	(-435,199)	
(2001-2002)	-194	(-2.6e+43,88)		28	(-144400,370)	

Table 5.16: Differences (with 95% Confidence Intervals) between annual Regional BMS indices for the South Eastern Region. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of the differences between 1000 simulated values using the Variance-Covariance matrix. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices (rounded to the nearest butterfly)

Years	No. Sites	Ratio A. Mean	95% CI	*	Ratio G. Mean	95% CI	*
(1983-1984)	8	2.761	(0.27,7.63)		1.923	(0.27,5.99)	
(1984-1985)	9	0.352	(0.09,1.56)		0.465	(0.12,1.72)	
(1985-1986)	7	1.999	(0.95,5.86)		1.309	(0.42,8.11)	
(1986-1987)	8	1.007	(0.16,8.35)		0.482	(0.11,7.35)	
(1987-1988)	9	0.913	(0.27,8.94)		0.512	(0.12,5.81)	
(1988-1989)	9	1.848	(0.36,5.76)		1.147	(0.09,5.72)	
(1989-1990)	15	0.907	(0.45,8.02)		1.671	(0.65,8.86)	
(1990-1991)	16	0.561	(0.17,1.56)		0.440	(0.12,1.08)	
(1991-1992)	16	0.804	(0.33,2.78)		0.943	(0.41,2.59)	
(1992-1993)	14	0.442	(0.17,1.62)		0.356	(0.08,1.74)	
(1993-1994)	13	1.659	(0.41,6.08)		0.808	(0.21,3.52)	
(1994-1995)	12	1.143	(0.20,3.63)		0.102	(0,3.04)	
(1995-1996)	12	2.694	(0.88,9.10)		1.452	(0.34,2e+27)	
(1996-1997)	15	0.754	(0.42,1.31)		1.053	(0.43,3.01)	
(1997-1998)	18	0.831	(0.17,2.51)		1.878	(0.31,4.84)	
(1998-1999)	25	0.586	(0.21,1.68)		0.612	(0.23,1.36)	
(1999-2000)	25	0.428	(0.31,1.80)		0.308	(0.13,1.11)	
(2000-2001)	23	3.433	(0,13.57)		3.070	(0.74,17.52)	
(2001-2002)	13	3.962	(0.51,109.41)		2.087	(0.31,57.05)	

Table 5.17: Ratios (with 95% Confidence Intervals) between annual Regional BMS indices for the South Eastern Region. Confidence Intervals were calculated using the 2.5th and 97.5th percentiles of the ratios between 1000 bootstrapped indices. Regional Indices were calculated by the arithmetic and geometric means of the selected site level indices, predicted at standardised covariates.

confidence interval does not contain 1 in this case).

Graphs of these ratios, with 95% confidence intervals are in Figures 5.18 and 5.19

[3]For Figures 5.18 and 5.19, unbounded ratios are due to upper bounds which are too large.

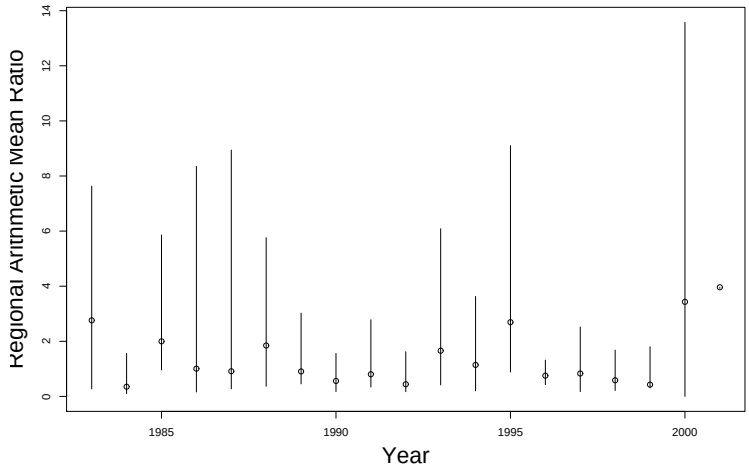


Figure 5.18: Arithmetic Ratios of Regional Indices between years for the South East Region, with 95% Confidence Intervals, calculated by bootstrapping sites. Ratios are calculated using only sites surveyed in both years.
3

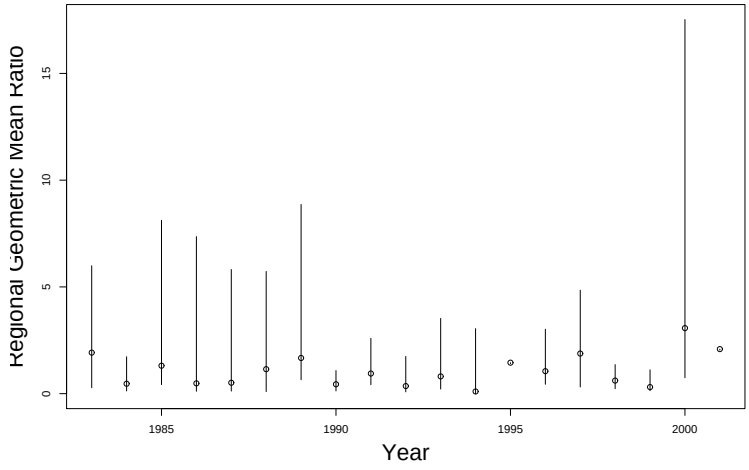


Figure 5.19: Geometric Ratios of Regional Indices between years for the South East Region, with 95% Confidence Intervals, calculated by bootstrapping sites. Ratios are calculated using only sites surveyed in both years.
3

5.7 Discussion

5.7.1 Use of a Regional Model

The development of a regional model, drawing strength and information from geographically close sites, seems to be well founded. Graphically, it is clear that sites within a region tend to share similar trends, though how dependable this is over larger areas is open to question. Whether the regional model selected is a GLM, GAM or GEE is also flexible - all can easily be implemented with our data.

Generally, all classes of model converged, and gave reasonable site-level predictions. It seems a sensible option to model at regional-level, rather than have individual models at site-level, with less data support.

5.7.2 Model and Covariate Selection

Model comparison and covariate selection for GEEs are challenging areas of research. QIC is available, which can be used to compare models, however, some questions still remain. For example, looking at Table 5.5, there is very little difference in QIC between a AR(1) GEE with linear altitude, and an exchangeable GEE without altitude. When correlation structure and covariate inclusion are both addressed simultaneously, it could pose problems for the modeler.

Any other, more widely accepted and used model selection criteria such as F-tests, depend on likelihood modelling techniques.

The fact that the AIC step-function and the QIC statistic choose the same model covariates for this data-set is encouraging. However, whether the same would occur when the correlation coefficient is stronger is currently unknown.

5.7.3 GAMM Results

The AIC for the mixed models are lower than that for the over-dispersed GLM (Table 5.2). This is probably due to the mixed models allowing for correlation,

which clearly exists in the data. However, whether the GAMM should be used, in-spite of the lower AIC values, is open to question. The partially fitted functions in Figure 5.7 show that the confidence intervals for a model with random BMSDay are a lot wider than those for random intercept or GLM.

5.7.4 Spatial Model Results

The huge reduction in AIC for a model with a common regional correlation coefficient indicates strongly that this model is to be preferred. The attempt to separate the correlation into separate geographical sub-regions (i.e., counties) does not seem worthwhile. The fitted curves for different covariates for the two models do not suggest any benefit in fitting a model with extra correlation parameters (Figures 5.9 to 5.11).

5.7.5 GEE Results

Once there is sufficient data, as there is in most cases, the GEE method with regression splines works well. The correlation coefficient ρ is significant for 4 out of the 9 regional models (Table 5.10). Figures 5.12 and 5.13 show a good smooth fit to the observed data for both the South East and Scotland regions for individual transects. Obviously, the fit is not exact, but the curve captures the flexibility and smoothness of the butterfly flight path over time.

5.7.6 Geometric versus Arithmetic Means

As expected, the regional indices calculated using the arithmetic means are consistently higher than those corresponding to the geometric means.

5.7.7 Site-Median versus Standard Covariates

Table 5.9 presents results for the South East calculated using both site-specific time-varying covariates and standardised covariates. Results do differ, though it should be remembered that these are point estimates, and it would be more

interesting to compare confidence intervals. Due to time and computing restraints, bootstrapping was only applied to data using prediction with standardised covariates. Some years' results (e.g. 2002) are very close and some (e.g. the arithmetic mean in 1977) diverge greatly. Possible reasons for this include having a less stable model due to insufficient data or having outliers in the data.

5.7.8 Comparing Indices over Time

Neither method (comparing bootstrapped differences and the chain-ratio method) indicate significant changes in butterfly indices over years (Tables 5.15 and 5.17). Figures 5.18 and 5.19 show ratios calculated using only sites surveyed in successive years. The reason for having no confidence intervals around the geometric ratios for 1995-1996 and 2001-2002, or the arithmetic ratio for 2001-2002, is that the confidence intervals are too wide (see Table 5.17) and so would distort the graphs.

There appears to be a significant decrease between the arithmetic means for the years 1997 and 1998 using the variance-covariance method of estimating confidence levels, though this does not appear using the bootstrapping method.

5.7.9 Precision Estimates: Bootstrapping versus Variance-covariance Method

Confidence intervals using the variance-covariance method (Chapter 3, Section 3.15.2) are much wider than those calculated using the bootstrapping method. This suggests that at least one of the two methods give biased variance estimates. It would be interesting to investigate this further.

Chapter 6

Discussion and Conclusion

There are many choices to be made when modelling data - choices between models, between covariates and between error structures. Every model is dependent on the data included in the analysis. Naturally, results will vary depending on decisions made and the question of interest being addressed.

There is much advantage to borrowing strength across geographically close sites and developing a regional model as often there is insufficient data to model individual sites.

Results presented in the previous chapter indicate very little practical difference between models. Fitted functions and measures of goodness-of-fit are virtually indistinguishable and established methods of model comparison such as AIC are not always useful when comparing different categories of model - e.g. GLM versus GEE. For these reasons, we needed to carefully examine and compare model assumptions and results.

The method of using regression splines within a GEE is one which addresses many of the issues regarding ecological data. It allows for a flexible relationship to be modelled between the response and any model covariates. It also allows for correlation within the counts, which is very important when considering variance around our estimates.

Obviously, issues still remain, even when using this reliable method. Model

selection should ideally be undertaken using the most suitable method for the question of interest - i.e., QIC should be used when using GEEs. Unfortunately, there is no fully automated software available yet for this, but it should be implemented where and when possible.

Many issues have been dealt with in this thesis - including over-dispersion, correlation, flexibility and precision. However, many issues still remain. Observer effect has been found to be an important covariate in some surveys (Link and Sauer [1997b]) - this issue has not been addressed for the BMS data-sets in this thesis. Issues of detection are not directly taken account of either, as we have no information, for example, on distance, which has been used to estimate detection and hence abundance in some surveys (Newson et al. [2005]).

Using the GEE method with regression splines seems to produce reliable estimates of relative abundance with corresponding confidence intervals calculated from bootstrapping sites, at least when the number of sites within a region is greater than around six.

Of course, it should be remembered that everything considered in this thesis is dependent on the fact that the sites in our data-sets were self-selected. Any future work considered by CEH or BC should attempt to address this issue.

Distinguishing long-term trend from short-term fluctuations is very difficult given the relatively short period of time for which we have sufficient data to analyse. The graphs with estimates of abundance with confidence intervals indicate some fluctuations, but very little in the way of long-term trend.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petran B.N. and F. Csàaki, editors, *International symposium on information theory*, pages 267–281. Akadèemiai Kiadi, 1973.

D.R. Anderson, K.P. Burnham, and G.C. White. Aic model selection in overdispersed capture-recapture data. *Ecology*, 75:1780–1793, 1994.

N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

J.A. Brown and M.C. Boyce. Line transect sampling of karner blue butterflies (lycaeides melissa samuelis). *Environmental and Ecological Statistics*, 5:81–91, 1998.

S.T. Buckland, D.R. Anderson, K.P. Burnham, J.L. Laake, D.L. Borchers, and L. Thomas. *Introduction to Distance Sampling.* Oxford University Press, 2001.

W.G. Cochran. *Sampling Techniques.* Wiley, 1963.

A.C. Davison and D.V. Hinkley. *Bootstrap Methods and their Application.* Cambridge Series in Statistical and Probabilistic Mathematics, 1997.

P.J. Diggle, P. Heagerty, K.-Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data.* Oxford Statistical Science Series, 2002.

R.M. Fewster, S.T. Buckland, and G.M. Siriwardena. Analysis of population trends for farmland birds using generalized additive models. *Ecology*, 81: 1970–1984, 2000.

G.M. Fitzmaurice. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, 51 (1):309–317, 1995.

J. Fox. *An R and S-Plus Companion to Applied Regression*. Sage Publications, 2002.

J.W. Hardin and J.M. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, 2001.

J.W. Hardin and J.M. Hilbe. *Generalized Estimating Equations*. Chapman & Hall, 2003.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

P.J. Heagerty and B.F. Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985, 2001.

N.J. Horton and S.R. Lipsitz. Review of software to fit generalized estimating equation regression models. *The American Statistician*, 53:160–169, 1999.

C.M. Hurvich and C.L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

K.L. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

W.A. Link and J.R. Sauer. New approaches to the analysis of population trends in land birds: Comment. *Ecology*, 78:2632–2634, 1997a.

W.A. Link and J.R. Sauer. Estimation of population trajectories from count data. *Biometrics*, 53:488–497, 1997b.

R.C. Littell, G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. *SAS System for Mixed Models*. SAS Publishing, 1996.

M.L. Mackenzie, C.R. Donovan, and B.H. McArdle. Regression spline mixed models: A forestry example. *Journal of Agricultural, Biological & Environmental Statistics*, 10 (4):394–410, 2005.

J.H. Marchant, R. Hudson, S.P. Carter, and P. Whittington. *Population Trends in British Breeding Birds*. British Trust for Ornithology, 1990.

B.W. McDonald. Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society. Series B.*, 55 (2):391–397, 1993.

A.D.R. McQuarrie and C. Tsai. *Regression and Time Series Model Selection*. World Scientific, 1998.

D. Moss and E. Pollard. Calculation of collated indices of abundance of butterflies, based on monitored sites. *Ecological Entomology*, 18:77–83, 1993.

M.D. Mountford. Estimation of population fluctuations with application to the common bird census. *Applied Statistics*, 2:135–143, 1982.

J.A. Nelder. Quasi-likelihood and pseudo-likelihod are not the same thing. *Journal of Applied Statistics*, 27:1007–1011, 2000.

S.E. Newson, R.J.W. Woodburn, D.G. Noble, S.R. Baillie, and R.D. Gregory. Evaluating the breeding bird survey for producing national population size and density estimates. *Bird Study*, 52:42–54, 2005.

P.M. North. A novel clustering method for estimating numbers of bird territories. *Applied Statistics*, 26:149–155, 1977.

W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.

J. Pannekoek and A. van Strien. Trim 3 manual - http://www.cbs.nl/en-gb/menu/themas/milieu-natuur-ruimte/natuur/methoden/trim/manual.htm. 2005.

E. Pollard. A method for assessing changes in the abundance of butterflies. *Biol. Conservation*, 12:115–134, 1977.

D. Raj. A note on the variance of the ratio estimate. *Journal of the American Statistical Association*, 59:895–898, 1964.

P. Rothery and D.B. Roy. Application of generalized additive models to butterfly transect count data. *Journal of Applied Statistics*, 28:897–909, 2001.

L. Thomas and K. Martin. The importance of analysis method for breeding bird survey population trend estimates. *Conservation Biology*, 10 (2):479–490, 1996.

G. Upton. A model for interyear change in the size of bird populations. *Biometrics*, 37:113–127, 1981.

A.J. van Strien, R. van De Pavert, D. Moss, T.J. Yates, C.A.M. van Swaay, and P. Vos. The statistical power of two butterfly monitoring schemes to detect trends. *Journal of Applied Ecology*, 34:817–828, 1997.

C.A.M. van Swaay, C.L. Plate, and A.J. van Strien. Monitoring butterflies in the netherlands: how to get unbiased indices. *Proc. Exper. Appl. Entomol.*, 13:21–28, 2002.

G. Verbeke and E. Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23:541–556, 1997.

R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447, 1974.

R. Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *J. Statist. Comput. Simul.*, 48:233–243, 1993.

S.L. Zeger, K.-Y. Liang, and P.S. Albert. Models for longitudinal data: A generalized estimationg equation approach. *Biometrics*, 44:1049–1060, 1988.

Appendix A

Code for GEE Model Selection and Prediction

```
bfly.func<- function(data) {

#################################################################


library(car)
library(splines)
library(gam)
library(geepack)
library(mvtnorm)
library(MASS)

####################0.  checking vifs#############################
```

```
data<-data[c(3,5,6,7,10,11,12,13,14,15,18,19)]
data<-na.omit(data)
data$sum<-ceiling(data$sum)

la<-log(data[,7])
############################################################
checkdata<-data[-c(1,7,9)]

check<-vif(lm(checkdata$sum~.,data=checkdata))

if (max(as.numeric(check))<5)
{
usedata<-checkdata
}
if (max(as.numeric(check))>5)
{
usedata<-checkdata[,-c(as.vector(which(check>5)))]
}
usedata<-cbind(usedata,data$BAPhab)
names(usedata)[ncol(usedata)]<-c("BAPhab")
check<-check[check<5]



###################1.   variable/model selection#######################

tid<-data$Tcode
resp<-data$sum
## making dataset for variable selection with all covariates, without la #####

vsdata<-usedata[-which(names(usedata)=="sum")]
```

```
xvarnom<-names(vsdata)

placeholder1<-which(names(vsdata)=="BMSDay")
placeholder2<-which(names(vsdata)=="BAPhab")
xvarnom<-xvarnom[-c(placeholder1,placeholder2)]

##### from here, need two functions, one with baphab and one without.

#define model object starting model

mod<-paste("gam(resp~ BMSDay+as.factor(BAPhab)",sep="")

for (i in (1:length(xvarnom))) {
mod<-paste(mod,"+",xvarnom[i],sep="") }
mod<-paste(mod,",family=poisson,data=vsdata)",sep="")

fit<-eval(parse(text=mod))



#define scope

scopeuse<-paste("BMSDay=~1+BMSDay+bs(BMSDay,knots=c(116,175)),
BAPhab=~1+as.factor(BAPhab)",sep="")
for (i in 1:length(xvarnom)) {
scopeuse<-paste(scopeuse,",",xvarnom[i],"=~1+",xvarnom[i],
"+bs(",xvarnom[i],",knots=mean(",xvarnom[i],"))",sep="") }

stepmod<-paste("step.gam(fit,scope=list(",scopeuse,"),data=usedata)",sep="")
stepmod2<-eval(parse(text=stepmod))
```

```
#### if rank is smaller than p, then remove some independent
variables ###

while(stepmod2$rank<length(stepmod2$coefficients))
{
mayberemove<-names(check[which(check==max(check))])
check<-check[-c(which(names(check)==mayberemove))]
try<-paste("update.formula(formula(stepmod2),~.-bs(",mayberemove,"
,knots=mean(",mayberemove,"))-",mayberemove,",data=usedata)",sep="")
#try<-paste("update.formula(formula(stepmod2),~.,data=usedata)",sep=""))
stepmod2<-eval(parse(text=try))
stepmod2<-gam(stepmod2,family=poisson,data=usedata,trace=FALSE) }

n<-nrow(usedata)

stepmod4<-step(gam(stepmod2,offset=la,family=poisson,data=usedata,trace=FALSE,k=log(n)))
form<-update.formula(formula(stepmod4),~.+offset(la))

stepmod5<-glm(form,data=usedata,family=poisson,trace=F)


usedata<-data.frame(Tcode=data$Tcode,usedata)


## gee part:
geemod<-geese(formula(stepmod5),id=usedata$Tcode,corstr="ar1",
family=poisson,data=usedata)

geeremove<-which(is.na(geemod$beta))
```

```
ifelse(length(geeremove)>0,
geemod<-geese(formula(stepmod5),id=usedata$Tcode,corstr="ind",
family=poisson,data=usedata), geemod<-geemod)


geeremove<-which(is.na(geemod$beta))

ifelse(length(geeremove)>0, geemod<-stepmod5, geemod<-geemod)



#########################################################################
###                  getting daily predictions                      ###
#########################################################################


## get med. for each tcode separately ie for each unique tcode pr
s'thing

usedata<-cbind(usedata,la=la)

td<-as.factor(usedata$Tcode) a<- split(usedata, td)

store<-matrix(NA,nrow=length(a),ncol=length(usedata)) for (i in
(1:length(a))) {
store[i,]<-(t(as.matrix(apply(as.data.frame(a[i]),2,median)))) }
store<-as.data.frame(store) names(store)<-names(usedata)

#########################################################
```

```
toremove<-rep(0,times=length(unique(usedata$BAPhab))) for(i in
1:length(unique(usedata$BAPhab))) {
name<-paste("as.factor(BAPhab)",unique(sort(usedata$BAPhab))[i],sep="")
toremove[i]<-max(0,which(names(stepmod5$coefficient)==name)) }
toremove<-toremove[-c(which(toremove==0))]


len<-length(20:230) keep<-matrix(0,nrow=length(a)*len,ncol=1)

dataset<-matrix(0,nrow=2,ncol=length(a))
dataset[1,]<-sort(unique(tid)) save<-matrix(0,nrow=len,ncol=1)

## get working for all 14 tcodes for(i in (1:length(a))) { print(i)
## get working for every bmsday in regional range ## for (j in
(1:len)) { store[i,]$BMSDay<-20+j-1
predx<-model.frame(delete.response(terms(stepmod5)),store[i,])
predx2<-cbind(rep(1,nrow(predx)),predx)
hab<-max(0,which(names(stepmod5$coefficient)==name))

name<-paste("as.factor(BAPhab)", store[i,]$BAPhab,sep="")
habkeep<-which(names(coefficients(stepmod5))==name)

#if habitat IS the baseline, remove all of toremove

ifelse(store[i,]$BAPhab==min(unique(usedata$BAPhab)),
todefremove<-toremove,
todefremove<-toremove[-c(which(toremove==habkeep))] )

ifelse(nrow(as.matrix(todefremove))>0,
coefs<-as.matrix(geemod$beta)[-c(todefremove)],
```

```
coefs<-as.matrix(geemod$beta) ) ifelse(
names(predx2)[3]=="as.factor(BAPhab)", predx2[3]<-1,
predx2[3]<-predx2[3] ) ifelse(
names(predx2)[3]=="as.factor(BAPhab)"&store[i,]$BAPhab==min(unique(usedata$BAPhab)),
predx2<-predx2[-c(3)], predx2<-predx2 )

coefs<-c(coefs,1) coefs<-as.matrix(coefs) predx2<-as.matrix(predx2)
predx2<-as.numeric(predx2) predx2<-t(predx2)
save[j]<-(exp(as.matrix(predx2)%*%coefs))
} keep[((1+len*(i-1)):(len*i))]<-save dataset[2,i]<-sum(save) }

print(dataset) dataset<-as.data.frame(dataset)



############ using arithmetic mean to get regional index
#####################
regionalindex<-(sum(dataset[2,]))/ncol(dataset)

return(data,dataset,regionalindex)
}







site.predictions<-bfly.func(data)
```

Appendix B

Habitat Classifications

Butterfly Conservation and the CEH use a system of coding corresponding to the EUNIS (European Nature Information System), involving 40 different habitat types, as in Table B.1. These are simplified to a broad system involving seven different habitat types for this thesis, as in Table B.2.

Code	BAP habitat	EUNIS Cross ref. code
1	Coastal habitats	A2
2	Coastal habitats	B1.4
3	Coastal habitats	B1.5
4	Coastal habitats	B1.6
5	Coastal habitats	B1.7
6	Coastal habitats	B1.8
7	Coastal habitats	B1.9
8	Coastal habitats	B2
9	Coastal habitats	B3
10	Fen, marsh and swamp	C3.1/C3.2/C3.3/C3.4
11	Fen, marsh and swamp	C3.5/C3.6/C3.7/C3.8
12	Bog	D1/D2
13	Fen, marsh and swamp	D4
14	Fen, marsh and swamp	D5
15	Calcareous grassland	E1.2
16	Acid grassland	E1.7
17	Neutral grassland	E2.1/E2.2
18	Improved grassland	E2.6
19	Fen, marsh and swamp	E3
20	Bracken	E5.3
21	Non-tall herb	F3.1
22	Non-Scrub	F3.1
23	Dwarf shrub heath	F4
24	Non-Scrub	F9
25	Boundary and linear features	FA
26	Broadleaved, mixed and yew woodland	G1
27	Coniferous woodland	G3
28	Broadleaved, mixed and yew woodland	G4
29	Non-lines of trees & parkland	G5.1
30	Broadleaved, mixed and yew woodland	G5.2
31	Broadleaved, mixed and yew woodland	G5.6/G5.7/G5.8
32	Boundary and linear features	E5.2
33	Arable and horticultural	G1.D
34	Inland rock	H
35	Arable and horticultural	I1.1
36	Arable and horticultural	I1.2
37	Arable and horticultural	I1.3
38	Arable and horticultural	I1.5
39	Built-up areas and gardens	I2
40	Inland rock	J3/J4/J6

Table B.1: EUNIS habitat types.

Habitat code	Description
1	Native woodland
2	Plantation woodland
3	Calcareous grassland scrub
4	Coastal (dunes, cliffs, marshes, etc.)
5	Farmland
6	Fen, moss, heathland
7	Various (parkland, mixed, upland, etc.)

Table B.2: BMS broad habitat types.

Appendix C

SAS Code

C.1 Regression Splines

```
proc transreg data=shse02 noprint; model identity(sum)=
    bspline(BMSDay/knots= 116,175)
    bspline(Temp/knots= 20)
    bspline(sun/knots=6)
    bspline(wind/knots= 1)
    bspline(north1/knots= 1071)
    bspline(east1/knots=5405)
    bspline(alt/knots=100)
    bspline(starttime/knots=150);
    id tcode region county BAPhab area Day Month
                Year  Transect_name la sun  bmsweek;
    output out=bspline;
run;

proc sort data=bspline; by tcode; run;
```

C.2 Mixed Model Code

```
%glimmix(data=bspline,
procopt=covtest method=ML,
stmts=%str(
    class Tcode BAPhab County;
    model sum=   bmsday_1 bmsday_2 bmsday_3 bmsday_4 bmsday_5
            sun_1 sun_2 sun_3 sun_4
            temp_1 temp_2 temp_3 temp_4
            wind_1 wind_2 wind_3 wind_4
            east1_1 east1_2 east1_3 east1_4
            north1_1 north1_2 north1_3 north1_4
            alt_1 alt_2 alt_3 alt_4
            starttime
        / solution covb;
random int; repeated /type=ar(1) subject=tcode ; ods output
covb=covb ConvergenceStatus=ConvergenceStatus; ), error=poisson,
link=log,offset=la, maxit=1000, options=type3) ;
```

C.3 Spatial Code

```
    %glimmix(data=bspline,
procopt=covtest method=ML ,
stmts=%str(
    class Tcode  County baphab;
    model sum=
            bmsday_1 bmsday_2 bmsday_3 bmsday_4 bmsday_5
            temp_1 temp_2 temp_3 temp_4
            sun_1 sun_2 sun_3 sun_4
            wind_1 wind_2 wind_3 wind_4
            east1_1 east1_2 east1_3 east1_4
```

```
            north1_1 north1_2 north1_3 north1_4
            alt_1 alt_2 alt_3 alt_4
            starttime
        / solution covb;
repeated /group=County type=ar(1) subject=tcode  rcorr; ods output
ConvergenceStatus=ConvergenceStatus covb=covb; ), error=poisson,
link=log,offset=la, maxit=1000 , options=type3) ;
```

Appendix D

Code for the Variance-Covariance Method

```
inner.func<- function(k)
{
bs.dat<-matrix(0,nrow=length(1:k),ncol=length(a))
bs.dat<-(as.data.frame(bs.dat))
for (k in (1:k))
{
print(paste("Simulation number =",k),sep="")
coefs<-c(as.vector(rmvnorm(1,est,covb)),1)
len<-length(20:220)
keep<-matrix(0,nrow=length(a)*len,ncol=1)
b.data<-matrix(0,nrow=2,ncol=length(a))
b.data[1,]<-sort(unique(tid))
## get working for all 14 tcodes
for(i in (1:length(a)))
```

```
{
## get working for every bmsday in regional range ##
save<-matrix(0,nrow=len,ncol=1)
for (j in (1:len))
{
store[i,]$BMSDay<-(min(usedata$BMSDay)+j-1)
predx<-model.frame(delete.response(terms(stepmod5)),store[i,])
predx2<-cbind(rep(1,nrow(predx)),predx)
name<-paste("as.factor(BAPhab)",predx2[1,3],sep="")
hab<-max(0,which(names(stepmod5$coefficient)==name))
todefremove<-toremove[-which(toremove==hab)]
ifelse(nrow(as.matrix(todefremove))>0,
bs.coefs<-as.matrix(coefs[-c(todefremove)]),
bs.coefs<-as.matrix(coefs[-c(tomayberemove)]))
ifelse(nrow(bs.coefs)==0, bs.coefs<-as.matrix(coefs),
bs.coefs<-bs.coefs) bs.coefs<-as.matrix(bs.coefs)
ifelse(names(predx2)[3]=="as.factor(BAPhab)", predx2[3]<-1,
predx2[3]<-predx2[3]) predx2<-as.matrix(predx2)
predx2<-as.numeric(predx2) predx2<-t(predx2)
save[j]<-(exp(as.matrix(predx2)%*%bs.coefs))
#exp(as.matrix(predx2)%*%bs.coefs)
}
keep[((1+len*(i-1)):(len*i))]<-save b.data[2,i]<-sum(save)
}
b.data<-as.data.frame(b.data) bs.dat[k,]<-b.data[2,]
}
print("bs.dat1") print(bs.dat)
}
```

Appendix E

Code for Bootstrapping Sites

```
bootstrap.func<-function(dataset,data)
{

#dataset<-site.predictions$dataset
#data<-site.predictions$data



#pickout n sites from the n

nn<-ncol(dataset)
sites<-round(runif(nn,1,nn))

sites<-dataset[1,sites]
sites<-as.numeric(sites)
```

```
keepdata<-matrix(ncol=ncol(data),nrow=nrow(data)*2)
for(i in 1:length(sites))
{
keep<-data[which(data$Tcode==sites[i]),]
keep<-as.matrix(keep)
empty<-min(which(is.na(keepdata)))
keepdata[empty:((empty+nrow(keep))-1),]<-keep
}

keepdata<-na.omit(keepdata)
keepdata<-as.data.frame(keepdata)
names(keepdata)<-c(names(data))

################### do model selection on these nn sites - keepdata




kla<-log(keepdata[,7])
#################################################################
kcheckdata<-keepdata[-c(1,7,9)]

kcheck<-vif(lm(kcheckdata$sum~.,data=kcheckdata))

if (max(as.numeric(kcheck))<5)
{
kusedata<-kcheckdata
}
if (max(as.numeric(kcheck))>5)
{
```

```
kusedata<-kcheckdata[,-c(as.vector(which(kcheck>5)))]
}
kusedata<-cbind(kusedata,keepdata$BAPhab)
names(kusedata)[ncol(kusedata)]<-c("BAPhab")
kcheck<-kcheck[kcheck<5]


####################1.   variable/model selection#######################

ktid<-keepdata$Tcode
kresp<-keepdata$sum
## making dataset for variable selection with
#####all covariates, without la

kvsdata<-kusedata[-which(names(kusedata)=="sum")]

kxvarnom<-names(kvsdata)

kplaceholder1<-which(names(kvsdata)=="BMSDay")
kplaceholder2<-which(names(kvsdata)=="BAPhab")
kxvarnom<-kxvarnom[-c(kplaceholder1,kplaceholder2)]

##### from here, need two functions, one with baphab and one without.

#define model object starting model

ifelse(length(unique(kvsdata$BAPhab))==1,
kmod<-paste("gam(kresp~ BMSDay",sep=""),
kmod<-paste("gam(kresp~ BMSDay+as.factor(BAPhab)",sep=""))
```

```
for (i in (1:length(kxvarnom)))
{
kmod<-paste(kmod,"+",kxvarnom[i],sep="")
}
kmod<-paste(kmod,",family=poisson,data=kvsdata)",sep="")




kfit<-eval(parse(text=kmod))


#define scope

ifelse(length(unique(kvsdata$BAPhab))==1,
kscopeuse<-paste("BMSDay=~1+BMSDay+bs(BMSDay,knots=c(116,175)),",sep=""),
kscopeuse<-paste("BMSDay=~1+BMSDay+bs(BMSDay,knots=c(116,175)),
BAPhab=~1+as.factor(BAPhab)",sep="")
)




for (i in 1:length(kxvarnom))
{
kscopeuse<-paste(kscopeuse,",",kxvarnom[i],"=~1+",kxvarnom[i],
"+bs(",kxvarnom[i],",knots=mean(",kxvarnom[i],"))",sep="")
}
```

```
kstepmod<-paste("step.gam(kfit,scope=list(",kscopeuse,"),data=kusedata)",sep="")
kstepmod2<-eval(parse(text=kstepmod))

#### if rank is smaller than p, then remove some independent variables ###

while(kstepmod2$rank<length(kstepmod2$coefficients))
{
kmayberemove<-names(kcheck[which(kcheck==max(kcheck))])
kcheck<-kcheck[-c(which(names(kcheck)==kmayberemove))]
ktry<-paste("update.formula(formula(kstepmod2),~.-bs(",kmayberemove,",
knots=mean(",kmayberemove,"))-",kmayberemove,",data=kusedata)",sep="")

kstepmod2<-eval(parse(text=ktry))
kstepmod2<-gam(kstepmod2,family=poisson,data=kusedata,trace=FALSE)
}

kn<-nrow(kusedata)

kstepmod4<-step(gam(kstepmod2,offset=kla,family=poisson,data=kusedata,trace=FALSE,
k=log(kn)))
kform<-update.formula(formula(kstepmod4),~.+offset(kla))

kstepmod5<-glm(kform,data=kusedata,family=poisson,trace=F)



kusedata<-data.frame(Tcode=keepdata$Tcode,kusedata)

## gee part:
kgeemod<-geese(formula(kstepmod5),id=kusedata$Tcode,corstr="ar1",
family=poisson,data=kusedata)
```

```
kgeeremove<-which(is.na(kgeemod$beta))

ifelse(length(kgeeremove)>0,
kgeemod<-geese(formula(kstepmod5),id=kusedata$Tcode,corstr="ind",
family=poisson,data=kusedata),
kgeemod<-kgeemod)


kgeeremove<-which(is.na(kgeemod$beta))

ifelse(length(kgeeremove)>0,
kgeemod<-kstepmod5,
kgeemod<-kgeemod)


#########################################################################
###                  getting daily predictions                    ###
#########################################################################


## get med. for each tcode separately ie for each unique tcode pr s'thing

kusedata<-cbind(kusedata,kla=kla)

ktd<-as.factor(kusedata$Tcode)
ka<- split(kusedata, ktd)

kstore<-matrix(NA,nrow=length(ka),ncol=length(kusedata))
```

```
for (i in (1:length(ka)))
{
kstore[i,]<-(t(as.matrix(apply(as.data.frame(ka[i]),2,median))))
}
kstore<-as.data.frame(kstore)
names(kstore)<-names(kusedata)

#########################################################

ktoremove<-rep(0,times=length(unique(kusedata$BAPhab)))
for(i in 1:length(ktoremove))
{
kname<-paste("as.factor(BAPhab)",sort(unique(kusedata$BAPhab))[i],sep="")
ktoremove[i]<-max(0,which(names(kstepmod5$coefficient)==kname))
}
ktoremove<-ktoremove[-c(which(ktoremove==0))]



klen<-length(20:230)
kkeep<-matrix(0,nrow=length(ka)*klen,ncol=1)

kdataset<-matrix(0,nrow=2,ncol=length(ka))
kdataset[1,]<-sort(unique(ktid))
ksave<-matrix(0,nrow=klen,ncol=1)

## get working for all 14 tcodes
for(i in (1:length(ka)))
{
print(i)
```

```
## get working for every bmsday in regional range ##
for (j in (1:klen))
{
kstore[i,]$BMSDay<-20+j-1
kpredx<-model.frame(delete.response(terms(kstepmod5)),kstore[i,])
kpredx2<-cbind(rep(1,nrow(kpredx)),kpredx)
#kname<-paste("as.factor(BAPhab)",kpredx2[1,3],sep="")
khab<-max(0,which(names(kstepmod5$coefficient)==kname))

kname<-paste("as.factor(BAPhab)", kstore[i,]$BAPhab,sep="")
khabkeep<-which(names(coefficients(kstepmod5))==kname)

#if habitat IS the baseline, remove all of toremove

ifelse(kstore[i,]$BAPhab==min(unique(kusedata$BAPhab)),
ktodefremove<-ktoremove,
ktodefremove<-ktoremove[-c(which(ktoremove==khabkeep))]
)



ifelse(nrow(as.matrix(ktodefremove))>0,
kcoefs<-as.matrix(kgeemod$beta)[-c(ktodefremove)],
kcoefs<-as.matrix(kgeemod$beta)
)
ifelse(
names(kpredx2)[3]=="as.factor(BAPhab)",
kpredx2[3]<-1,
kpredx2[3]<-kpredx2[3]
)
ifelse(
```

```
names(kpredx2)[3]=="as.factor(BAPhab)"&kstore[i,]$BAPhab==min(unique(kusedata$BAPhab)),
kpredx2<-kpredx2[-c(3)],
kpredx2<-kpredx2
)









kcoefs<-c(kcoefs,1)
kcoefs<-as.matrix(kcoefs)
kpredx2<-as.matrix(kpredx2)
kpredx2<-as.numeric(kpredx2)
kpredx2<-t(kpredx2)
ksave[j]<-(exp(as.matrix(kpredx2)%*%kcoefs))
}
kkeep[((1+klen*(i-1)):(klen*i))]<-ksave
kdataset[2,i]<-sum(ksave)
}

print(kdataset)
kdataset<-as.data.frame(kdataset)

kkdataset<-matrix(ncol=length(sites),nrow=2)
kkdataset[1,]<-sites

for(i in 1:length(sites))
```

```
{
use<-which(kdataset[1,]==kkdataset[1,i])
#print(use)
kkdataset[2,i]<-kdataset[2,use]
}


kkdataset<-cbind(kkdataset,kregionalindex,kkregionalindex)


return(kkdataset)

}
```